Class 1 Introduction to Unsupervised Learning

Agenda

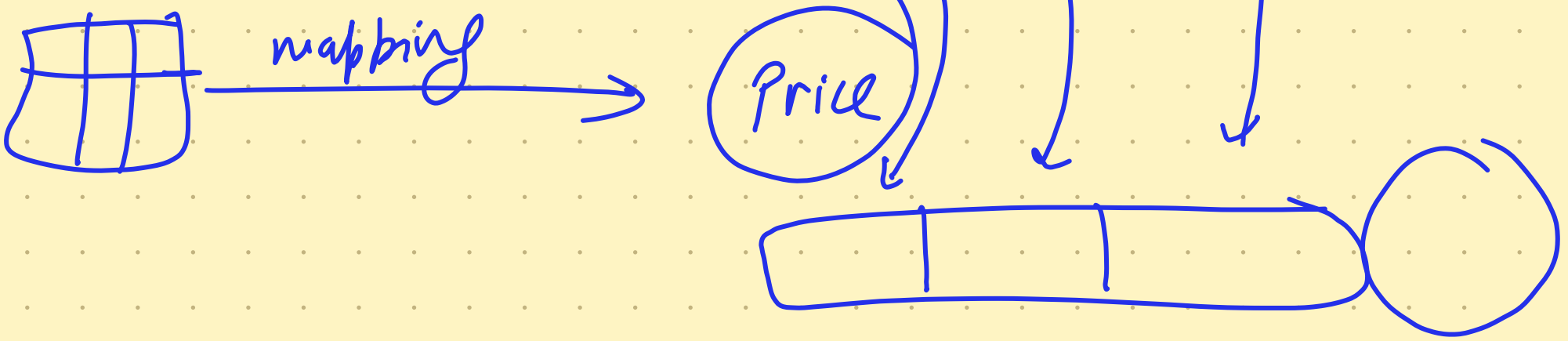
- ML setup
- Intro to Unsupervised Learning (Problem statement)
- Clustering (Intuition)
- Dunn Index
- K Means Algo
- LLoyd's Algo

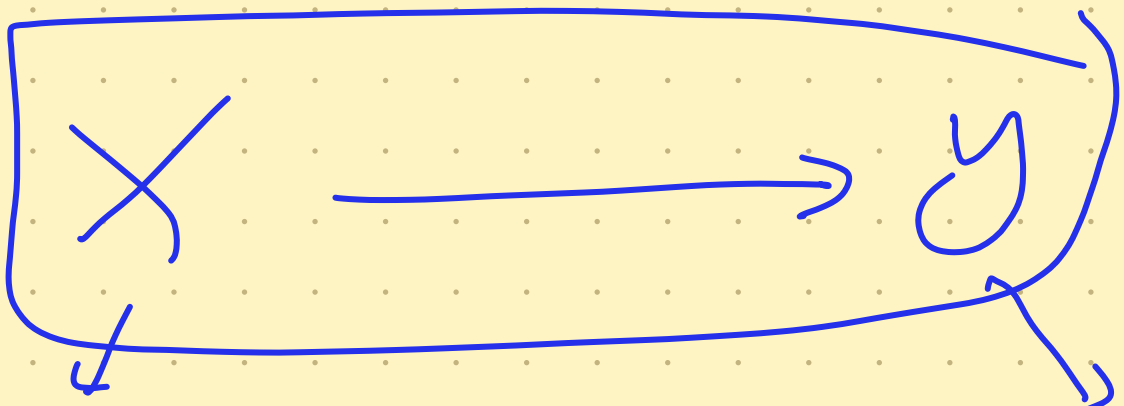
input → Train a model → output

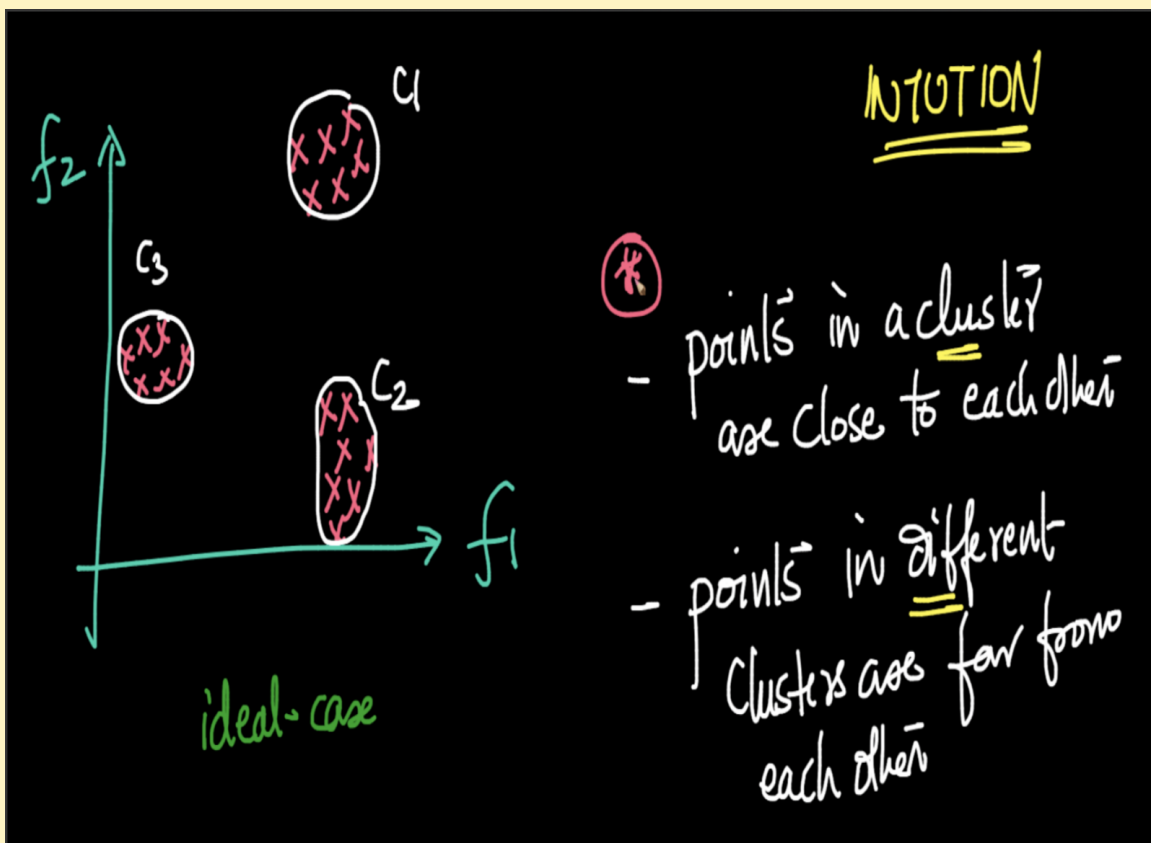
Data

Training Data

Area m^2	Bedrooms	Bathrooms	Price
200	3	2	$500,000
190	2	1	$450,000
230	3	3	$650,000
180	1	1	$400,000
210	2	2	$550,000







Amazon Data

	ID	n_clicks	n_visits	amount_spent	amount_discount	days_since_registra
0	1476	130	65	213.905831	31.600751	
1	1535	543	46	639.223004	5.689175	
2	1807	520	102	1157.402763	844.321606	
3	1727	702	83	1195.903634	850.041757	
4	1324	221	84	180.754616	64.283300	

X labels

supervised M.L

House features	Price

House	Area	No of bed room	

unsupervised

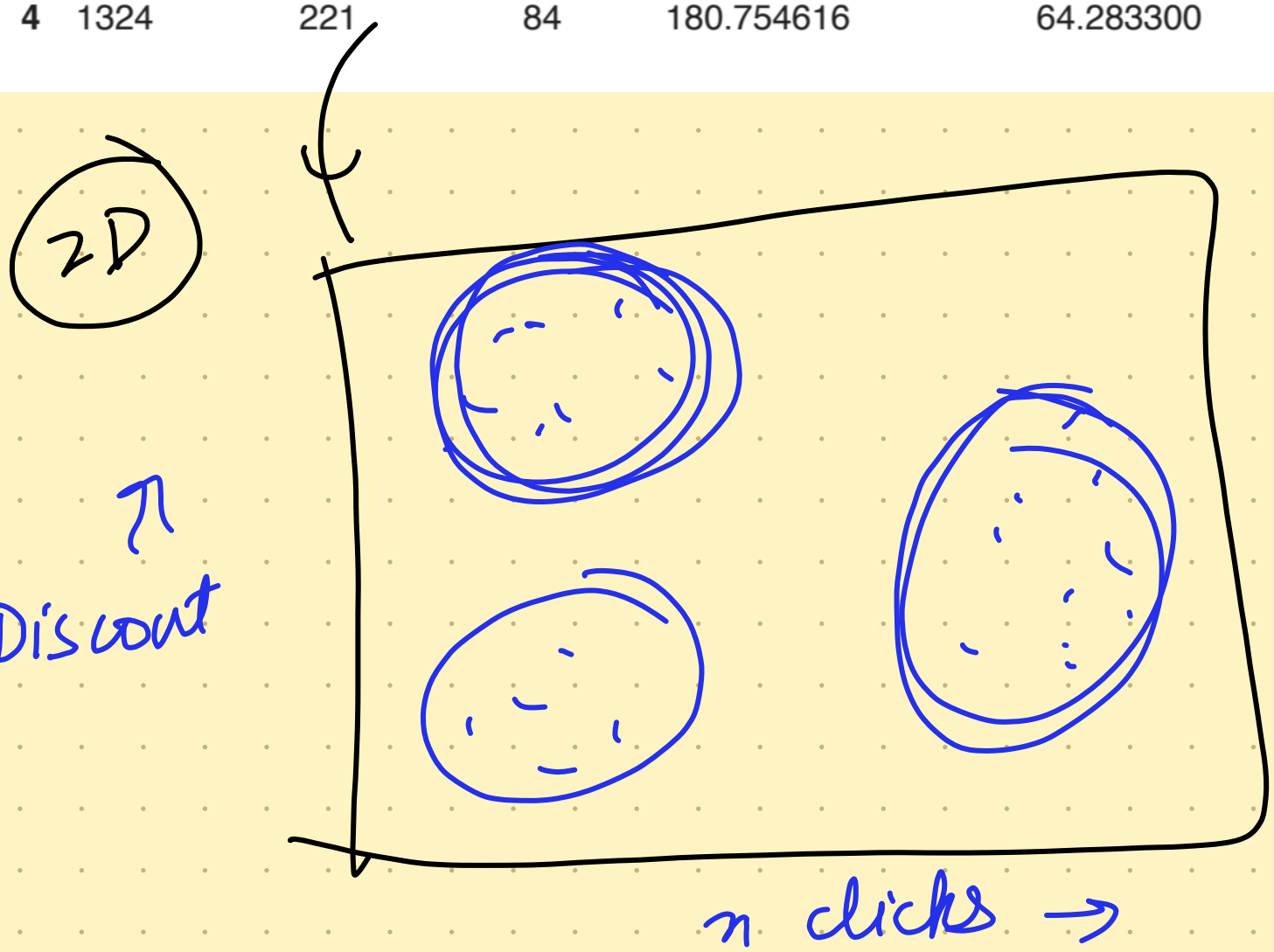
Supervised Maching Learning

- Regression ( )
- Classification

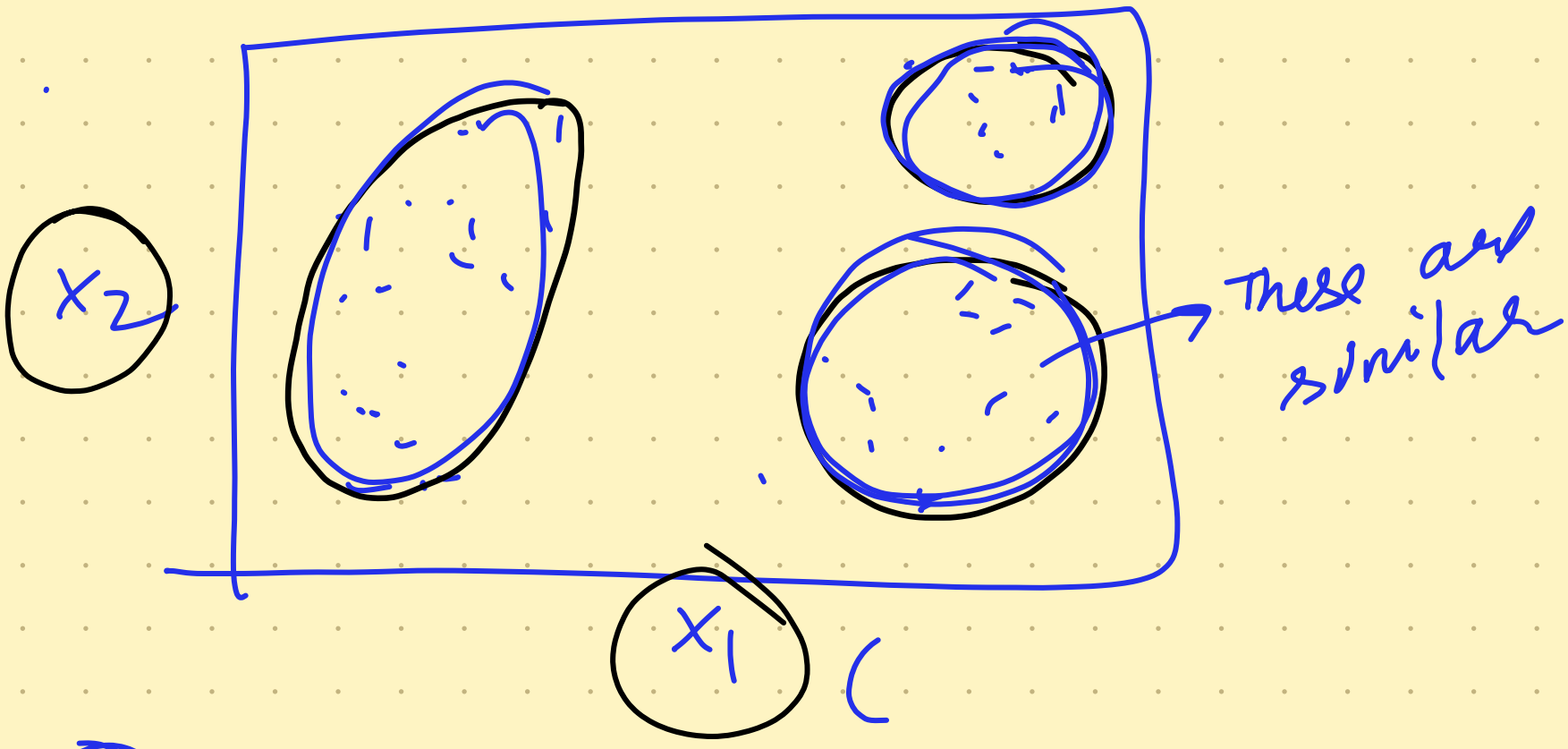
X and Y

x_1 x_2 ---	Y → Price → cat / dog 0 - 1

	ID	n_clicks	n_visits	amount_spent	amount_discount	days_since_registra
0	1476	130	65	213.905831	31.600751	
1	1535	543	46	639.223004	5.689175	
2	1807	520	102	1157.402763	844.321606	
3	1727	702	83	1195.903634	850.041757	
4	1324	221	84	180.754616	64.283300	

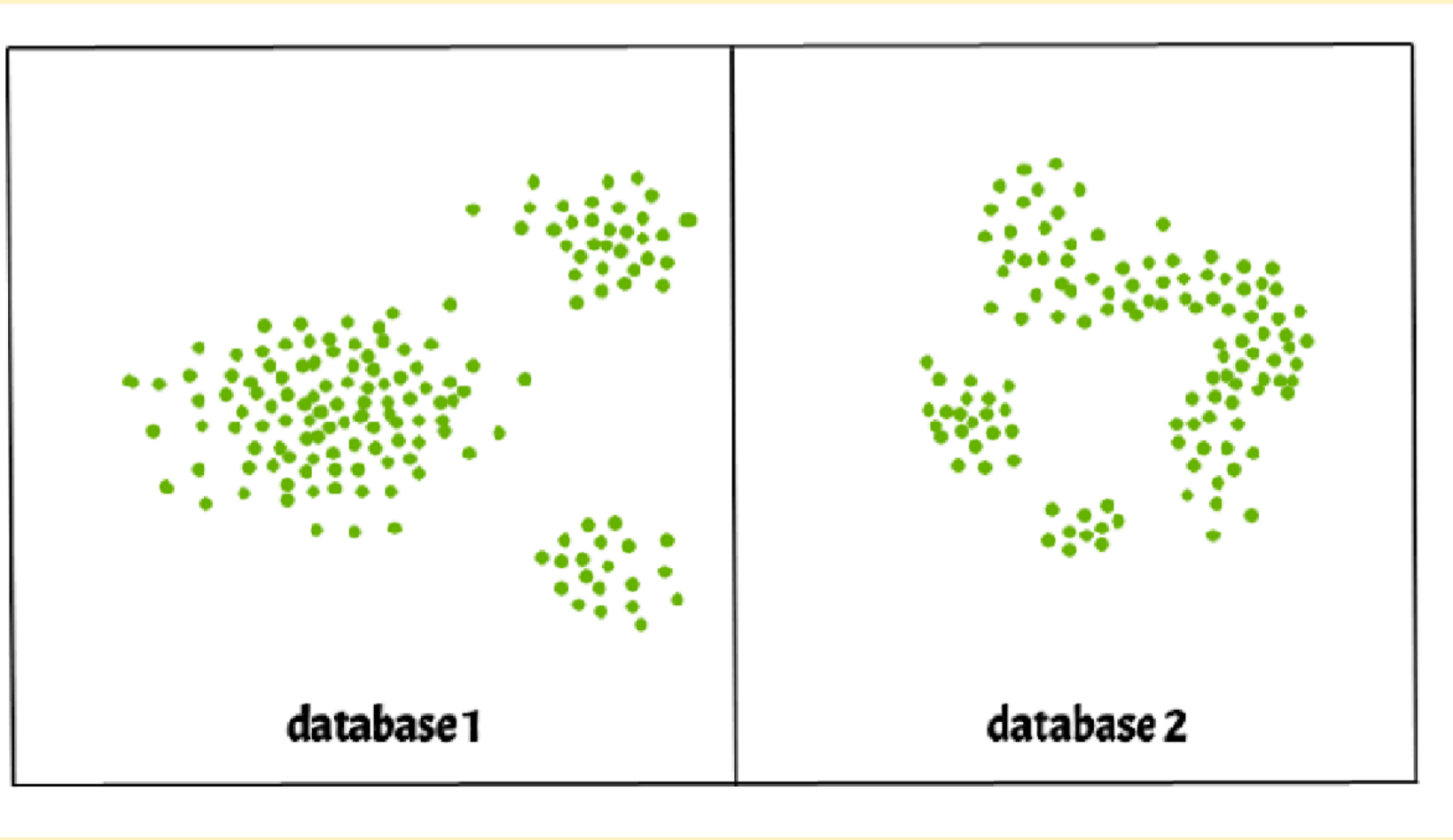


Unsupervised Learning →



y

Clustering [cricket Academy] → { 30,000 enrollment }

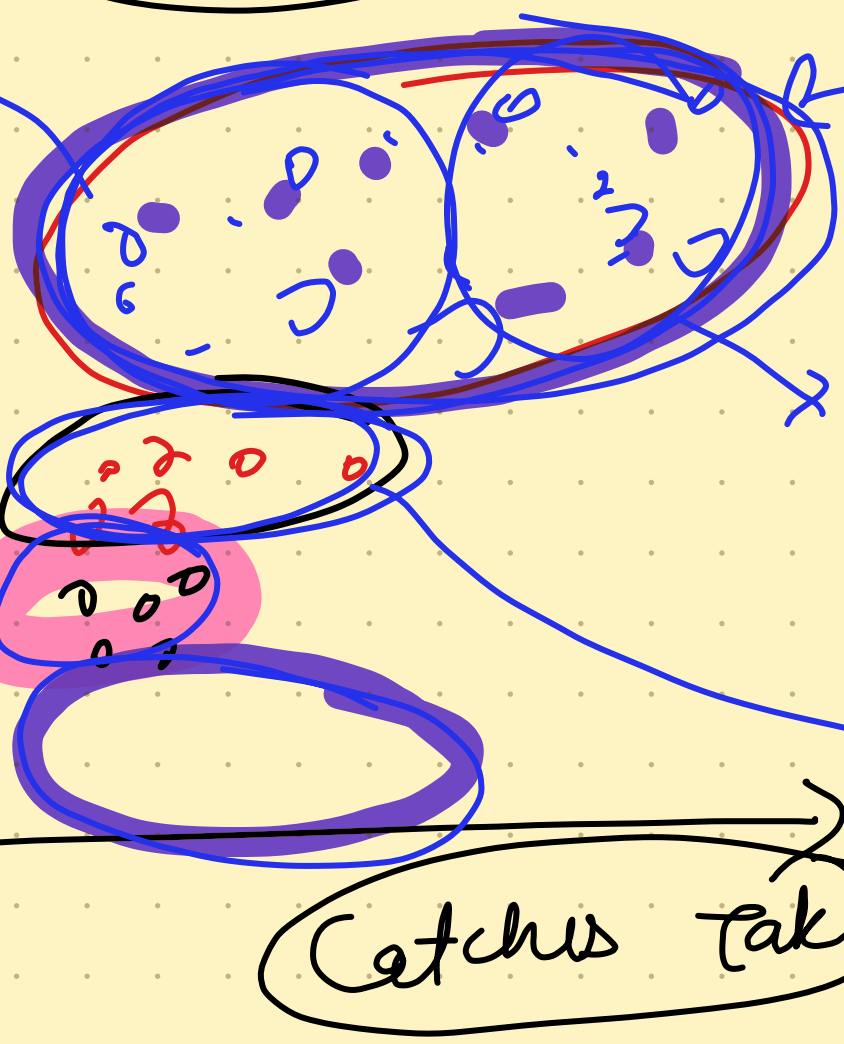


Stipend → Based on the performance of cricketers

Coach ⇒ Runs per game & wickets per game & catches per game

stipend 1

Runs Scored



Batsmen

similar to each other

stipend 2

Catches Taken →

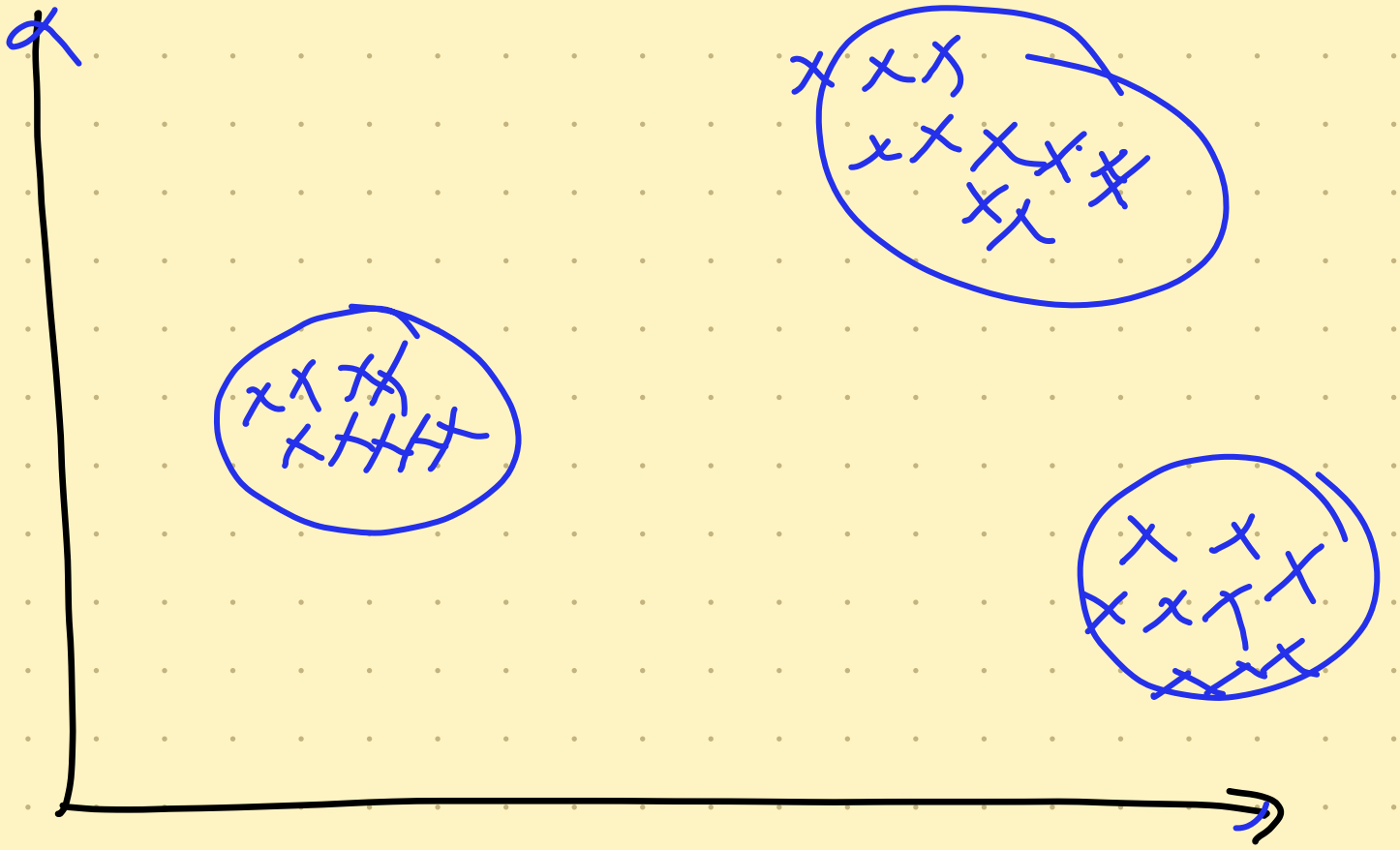
Wickets Taken

Bowlers

Catches Taken

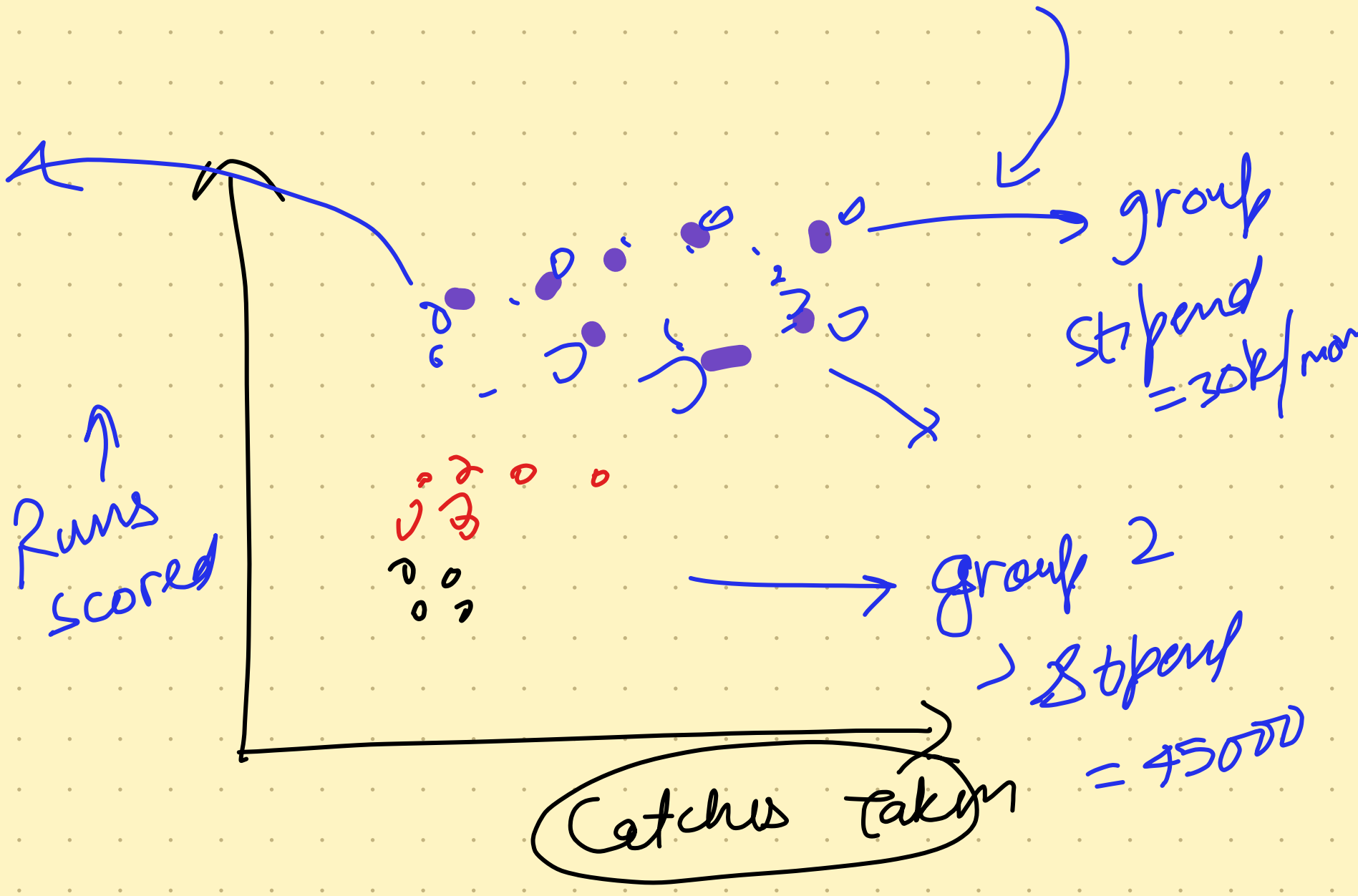
Player	Catches	Runs Scored
1		
2		
3		
⋮		
150		

Ideal case

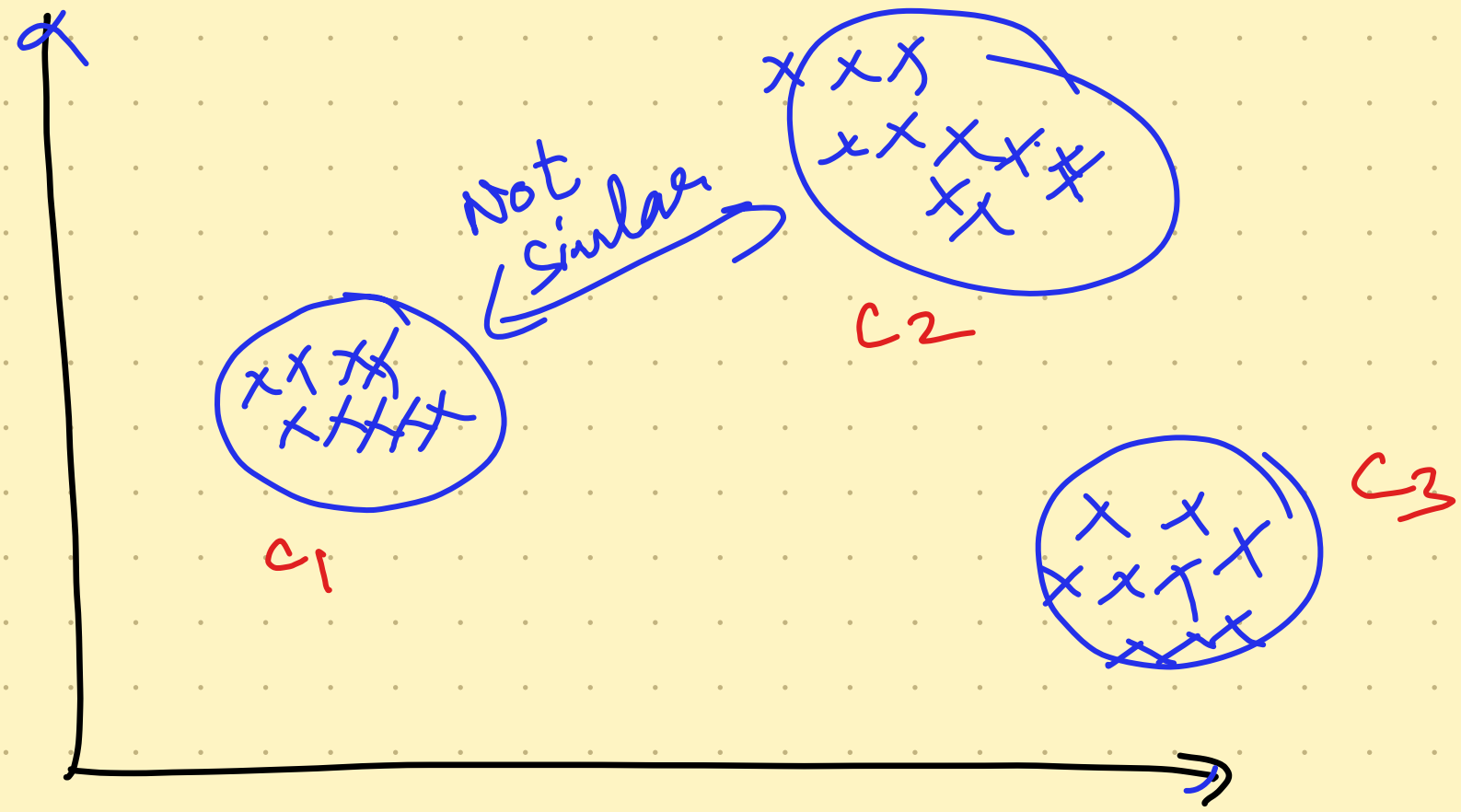


Intuition

Real world



Ideal case



Can we say (1) Points in a cluster are close to each other

(2) Points in different cluster are far from each other

wickets	Catches
3	5
2	6
3	6

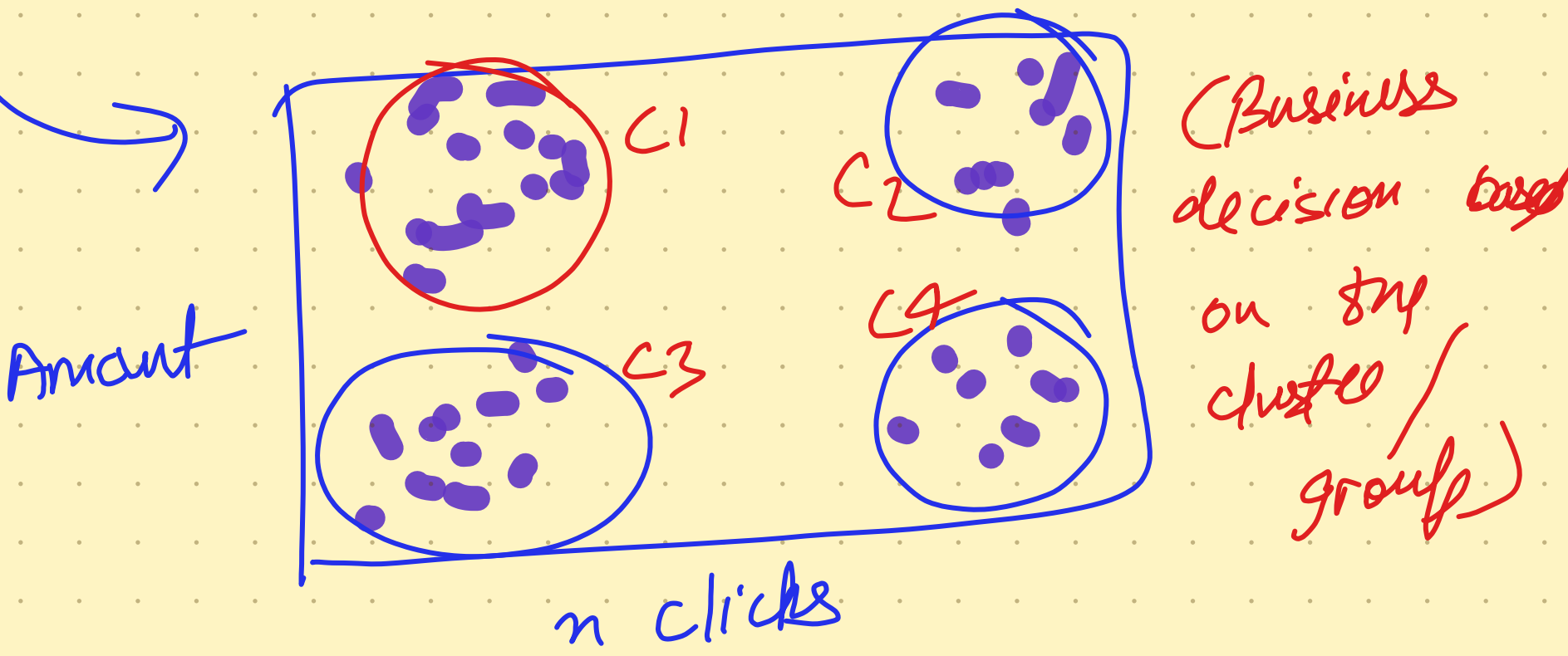
→ Plotted on paper

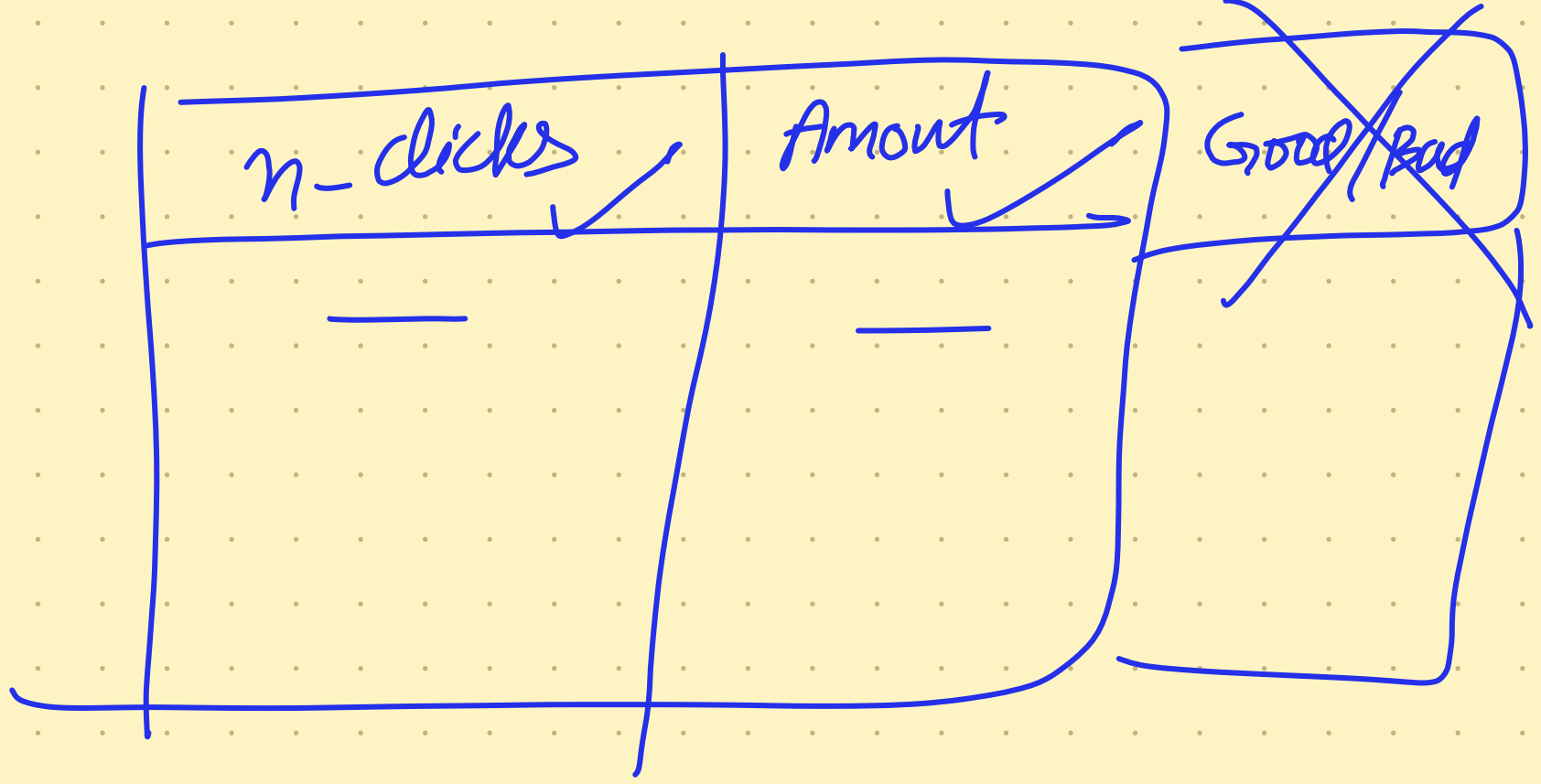
Q: what is the defin" of Similarity

→ close pts are similar
(closeness should make Business sense as well)

	ID	n_clicks	n_visits	amount_spent	amount_discount	days_since_registra
0	1476	130	65	213.905831	31.600751	
1	1535	543	46	639.223004	5.689175	
2	1807	520	102	1157.402763	844.321606	
3	1727	702	83	1195.903634	850.041757	
4	1324	221	84	180.754616	64.283300	

→ customers spending equal amt are similar

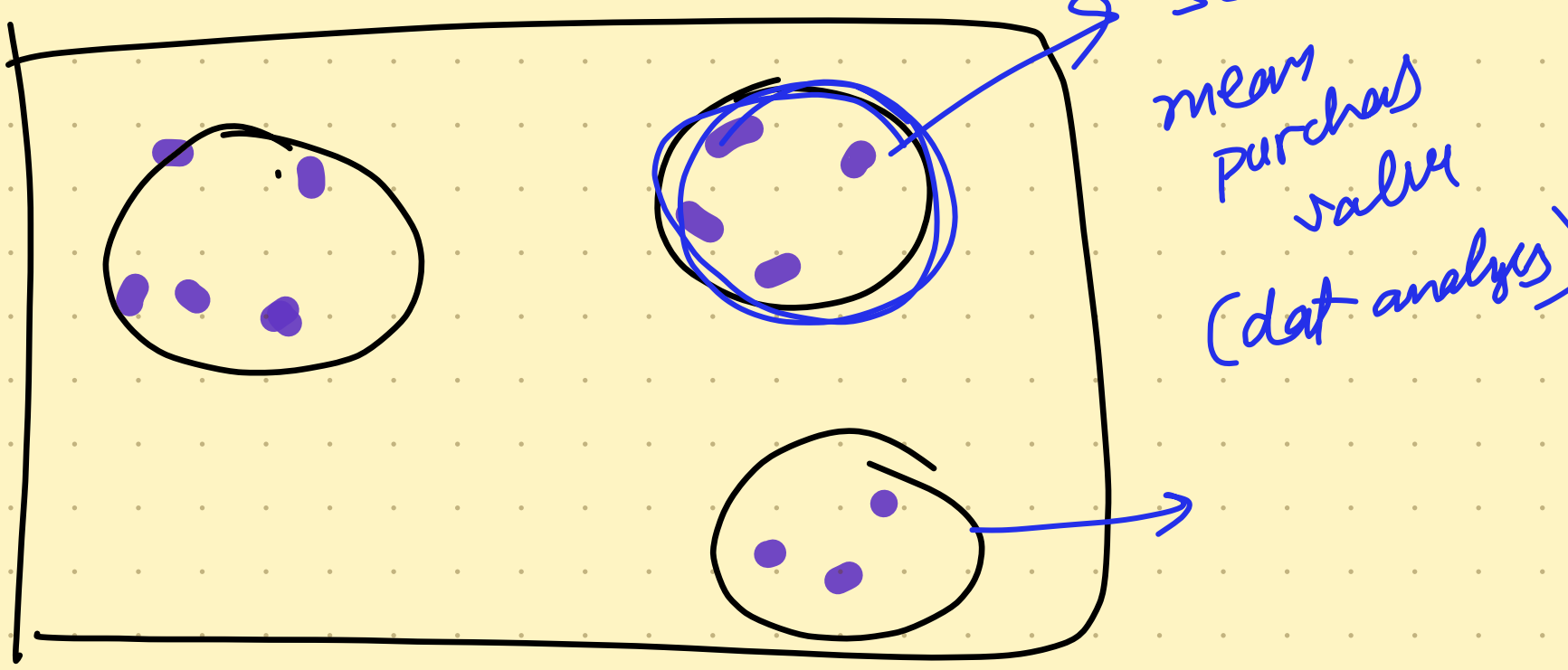
→ 

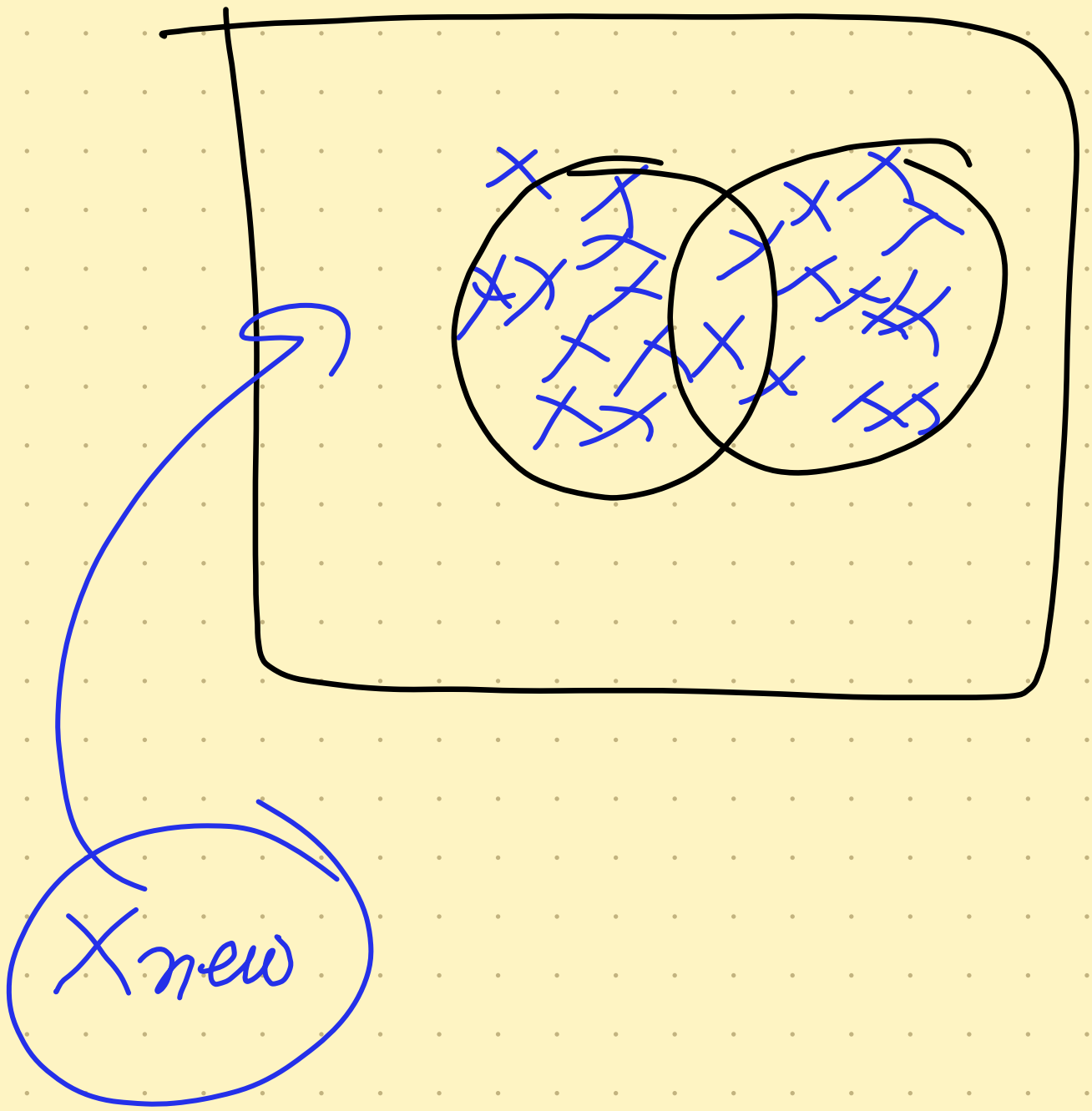


How to ensure clustering is good or Bad?

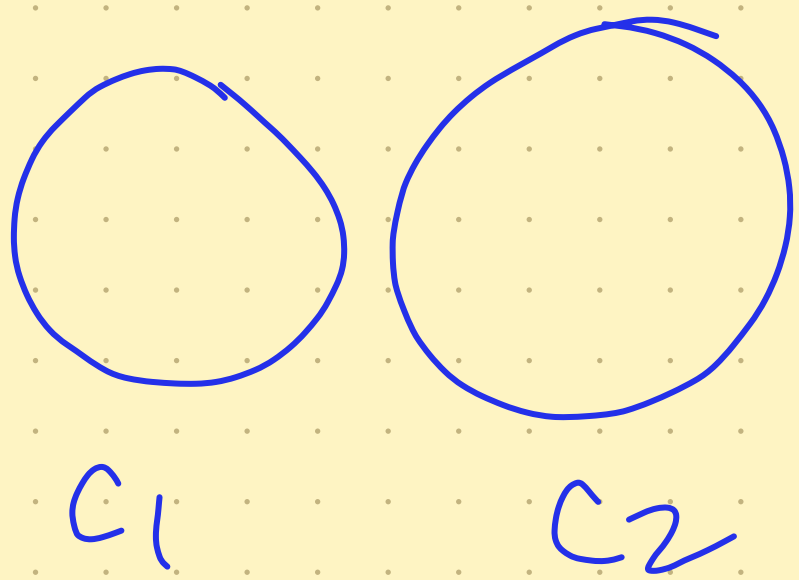
→ Ans. No one knows its good / Bad
(we don't have Y label)

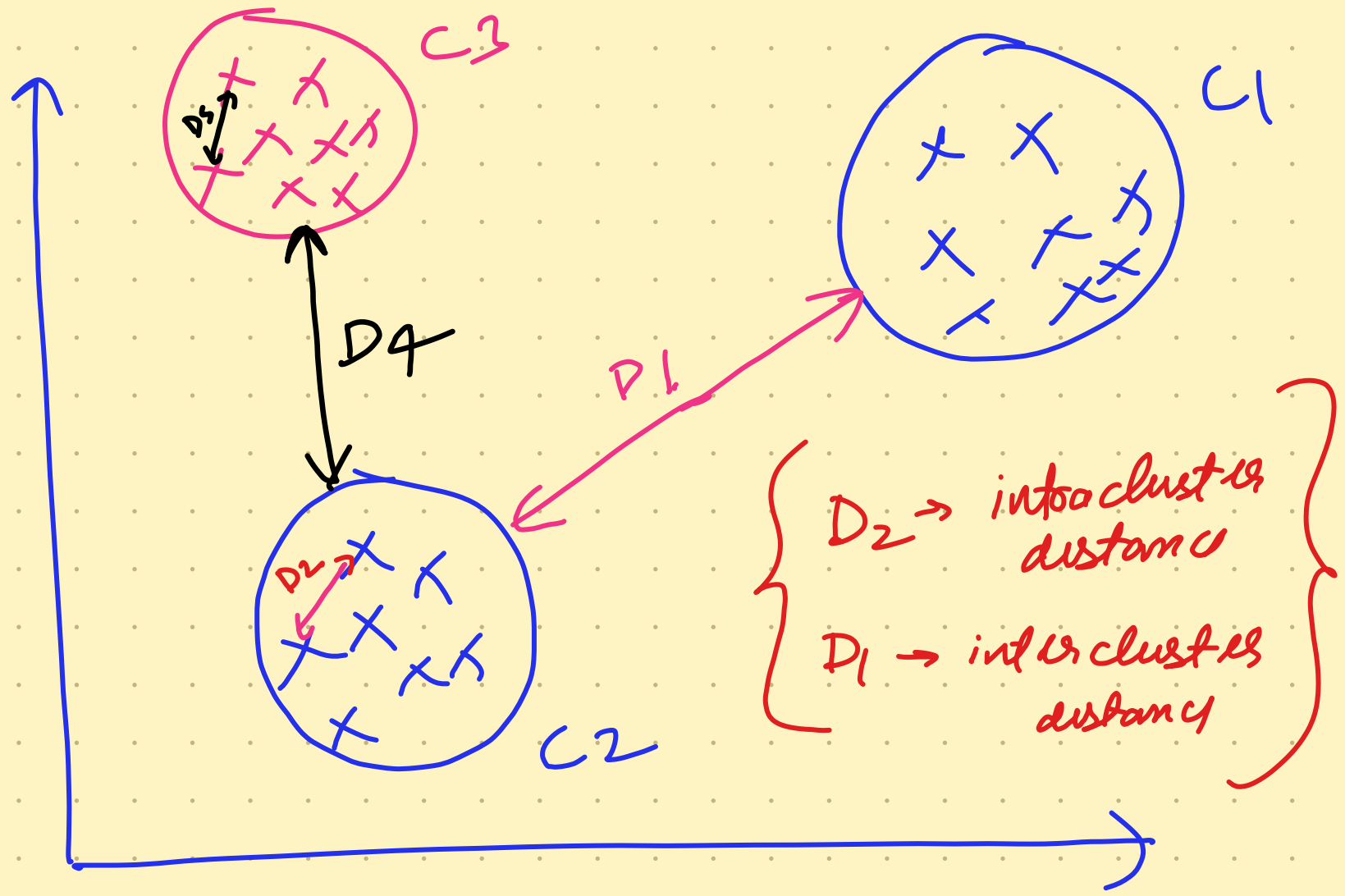
→ Business Sense





Terms and Glossary





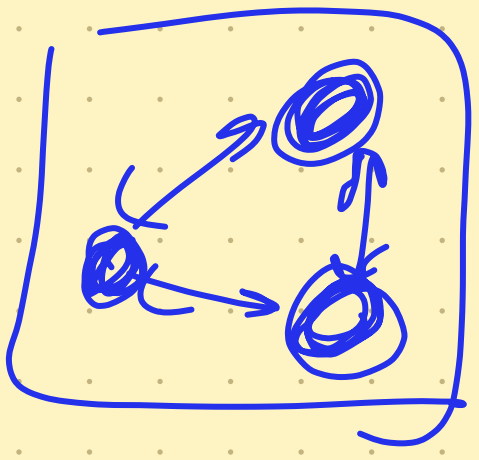
Inter College → B/w diff colleges

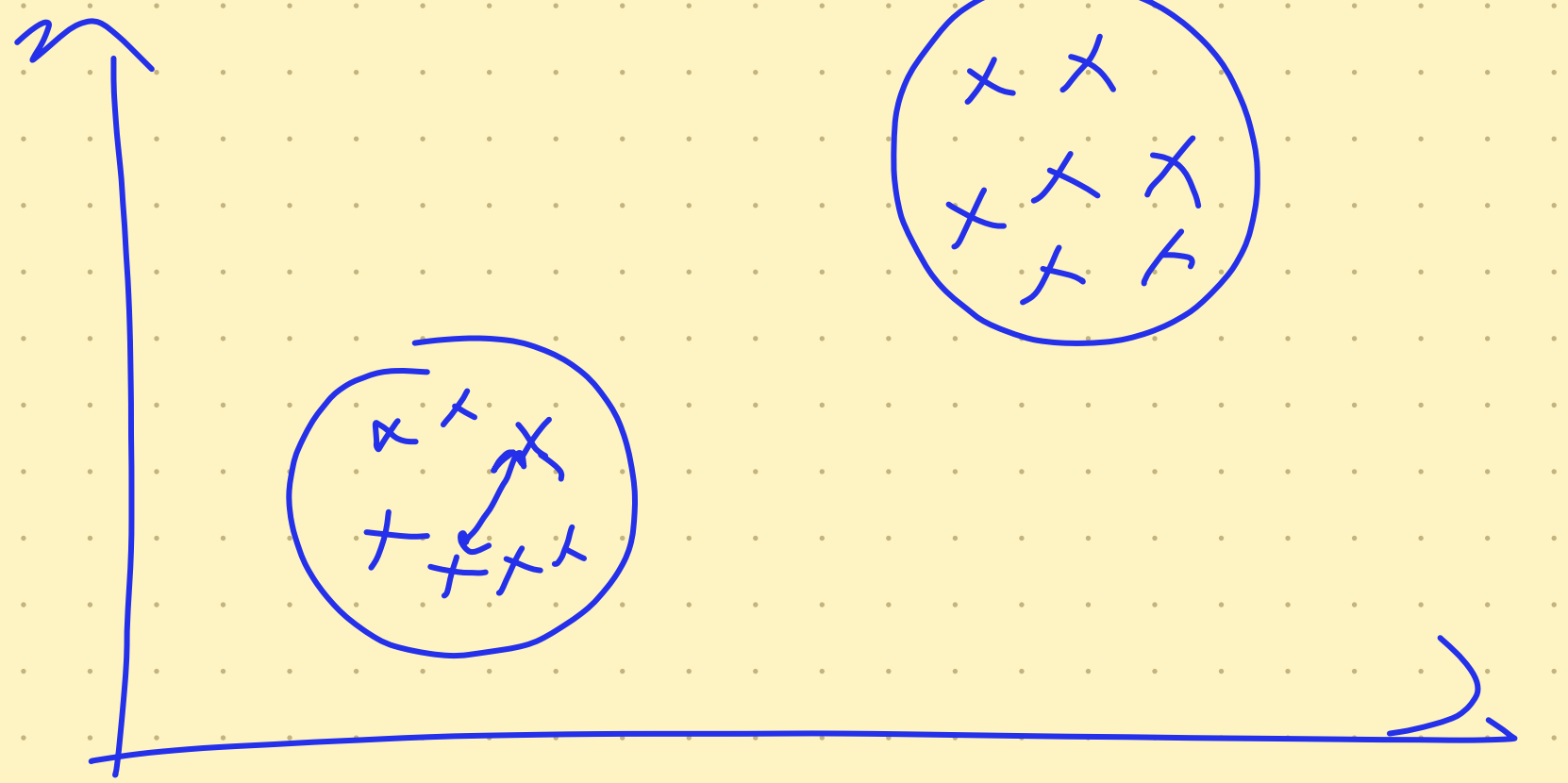
Intra College → within the same college

Distance

(maximise) ← Intra cluster

Intra cluster → (minimise)



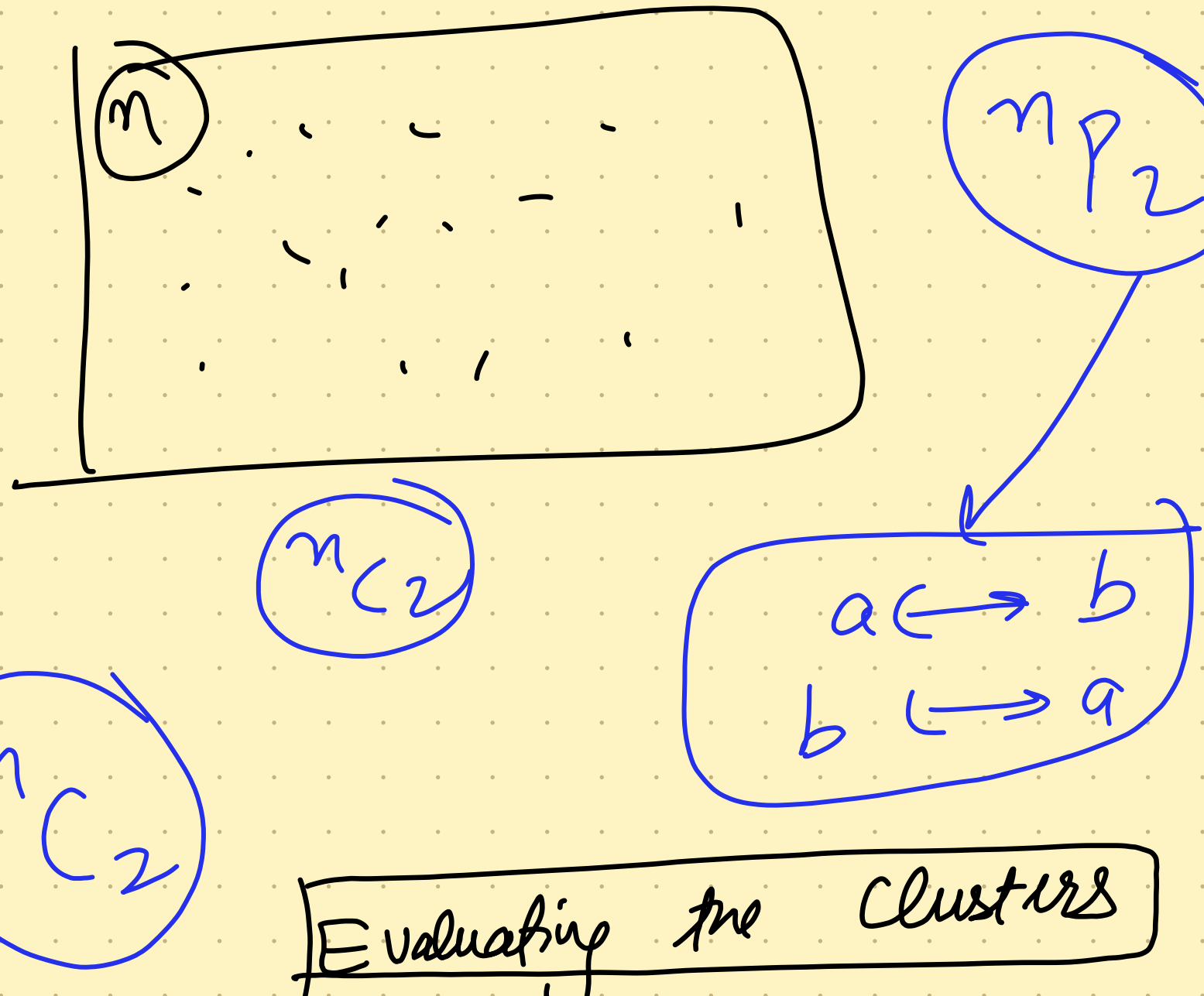


Practical

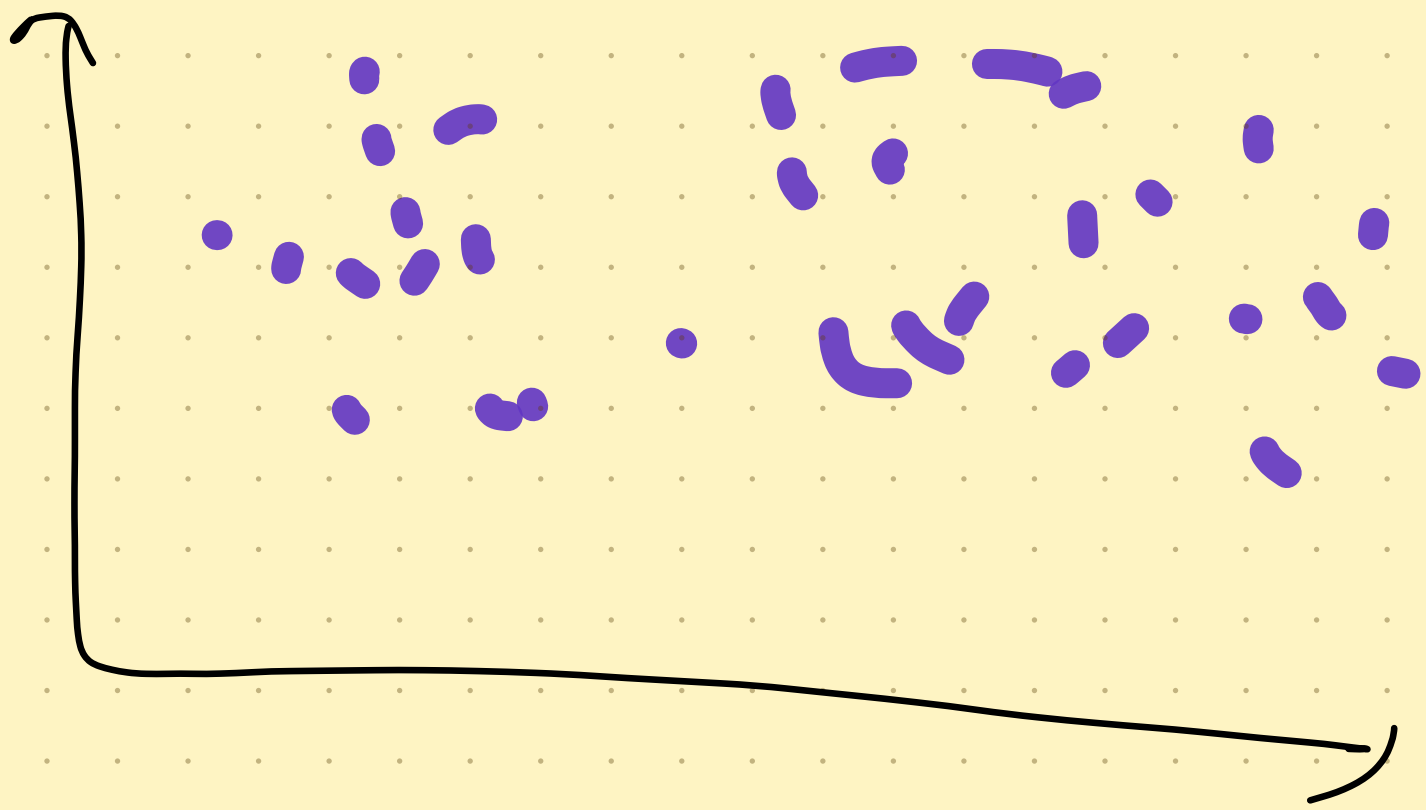
Distance Metric

- Euclidian ⇒ (low dim.)
- Manhattan ⇒ low to med dimension
- Cosine ⇒ high dimension

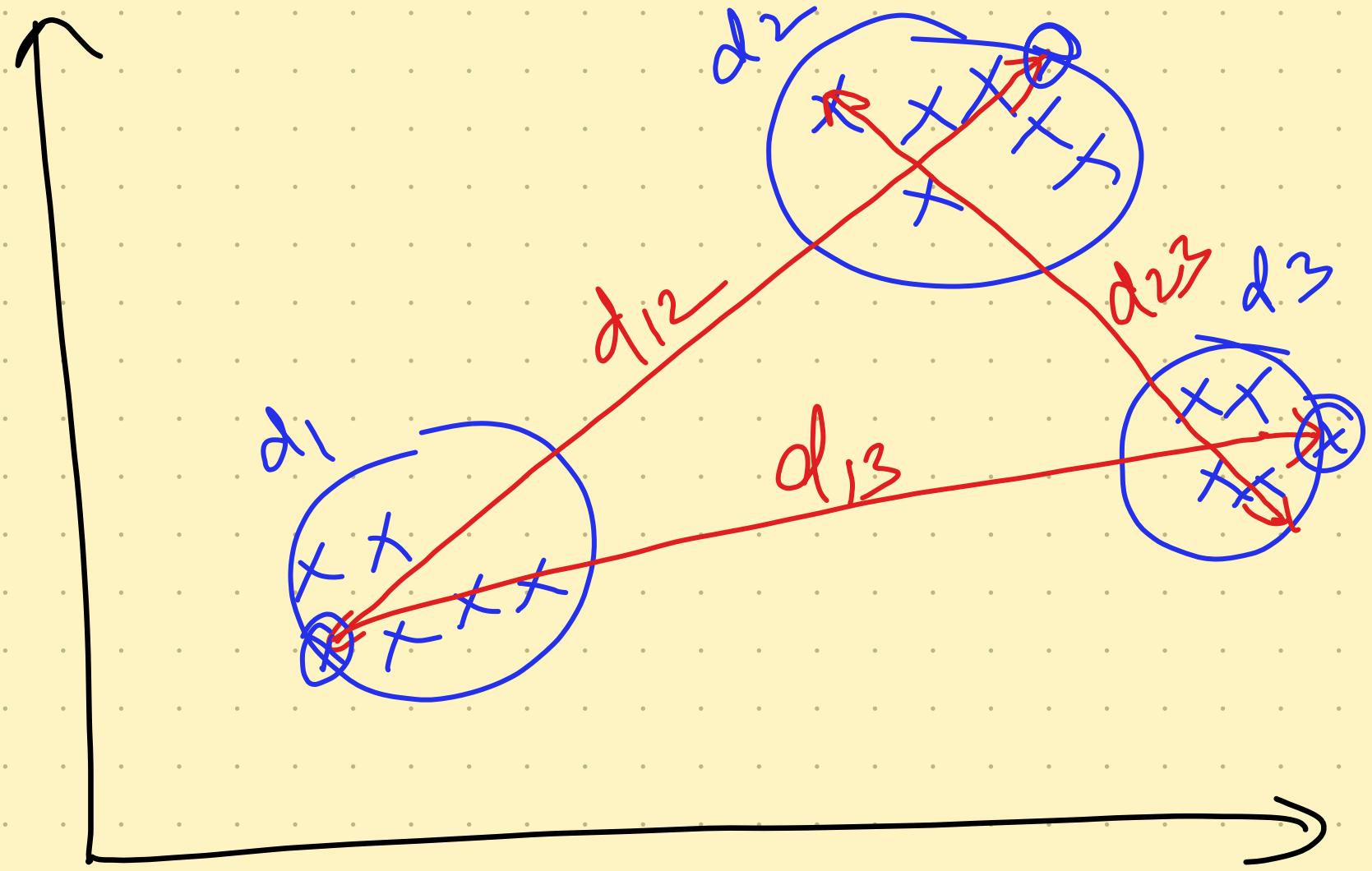
Euclidean ⇒ $\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$



Evaluating the clusters

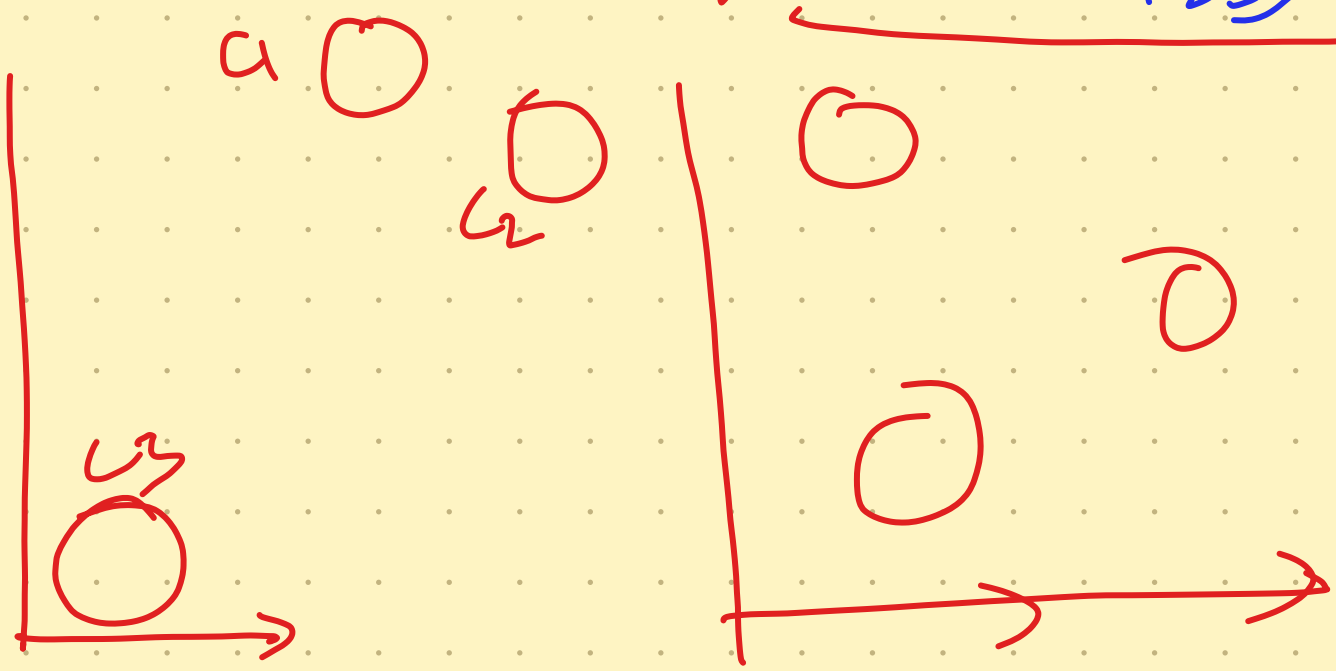


(#) [We still cannot evaluate how good the clusters are]

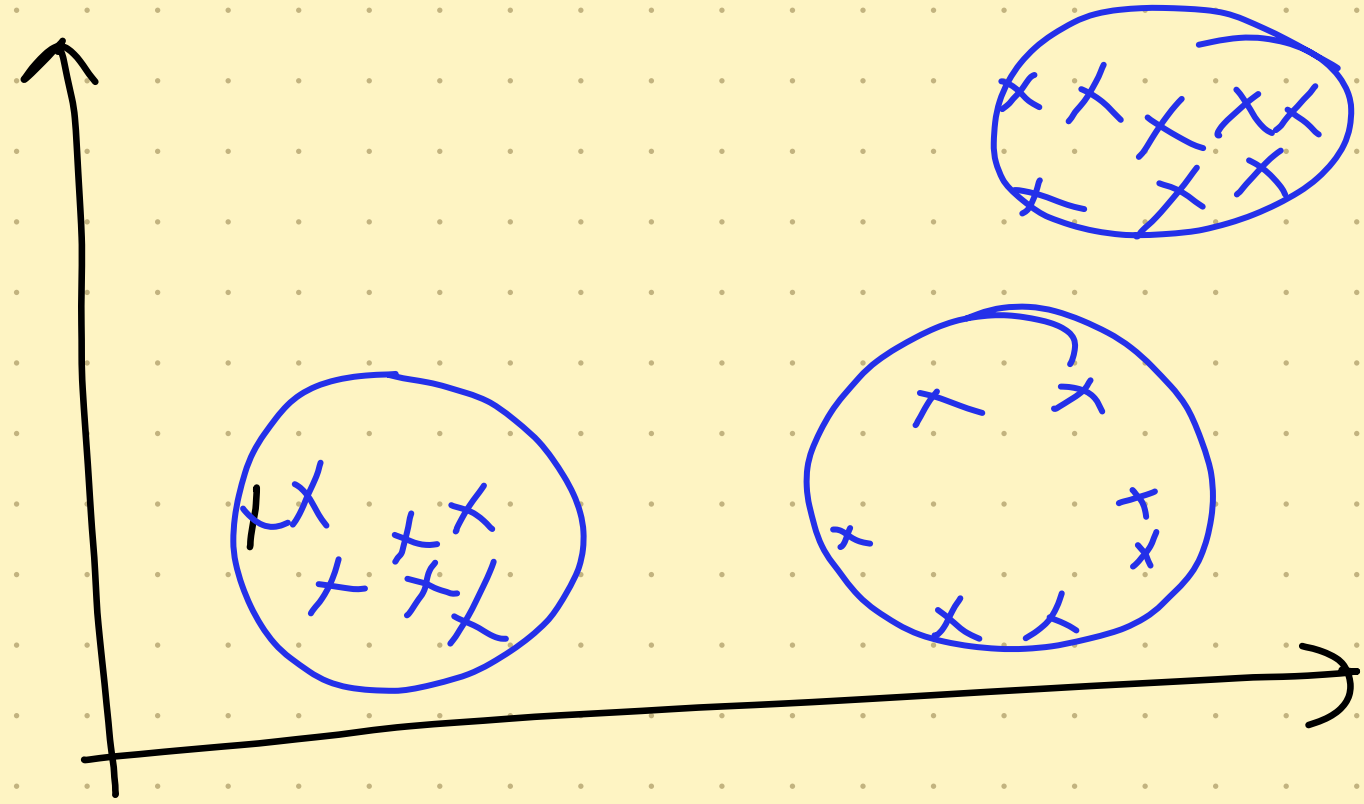


$d(i,j)$ = distance b/w farthest points in c_i and c_j [$c_i = c_1 / c_2 / c_3$]

min (d_{12}, d_{13}, d_{23}) // max (d_{12}, d_{13}, d_{23})

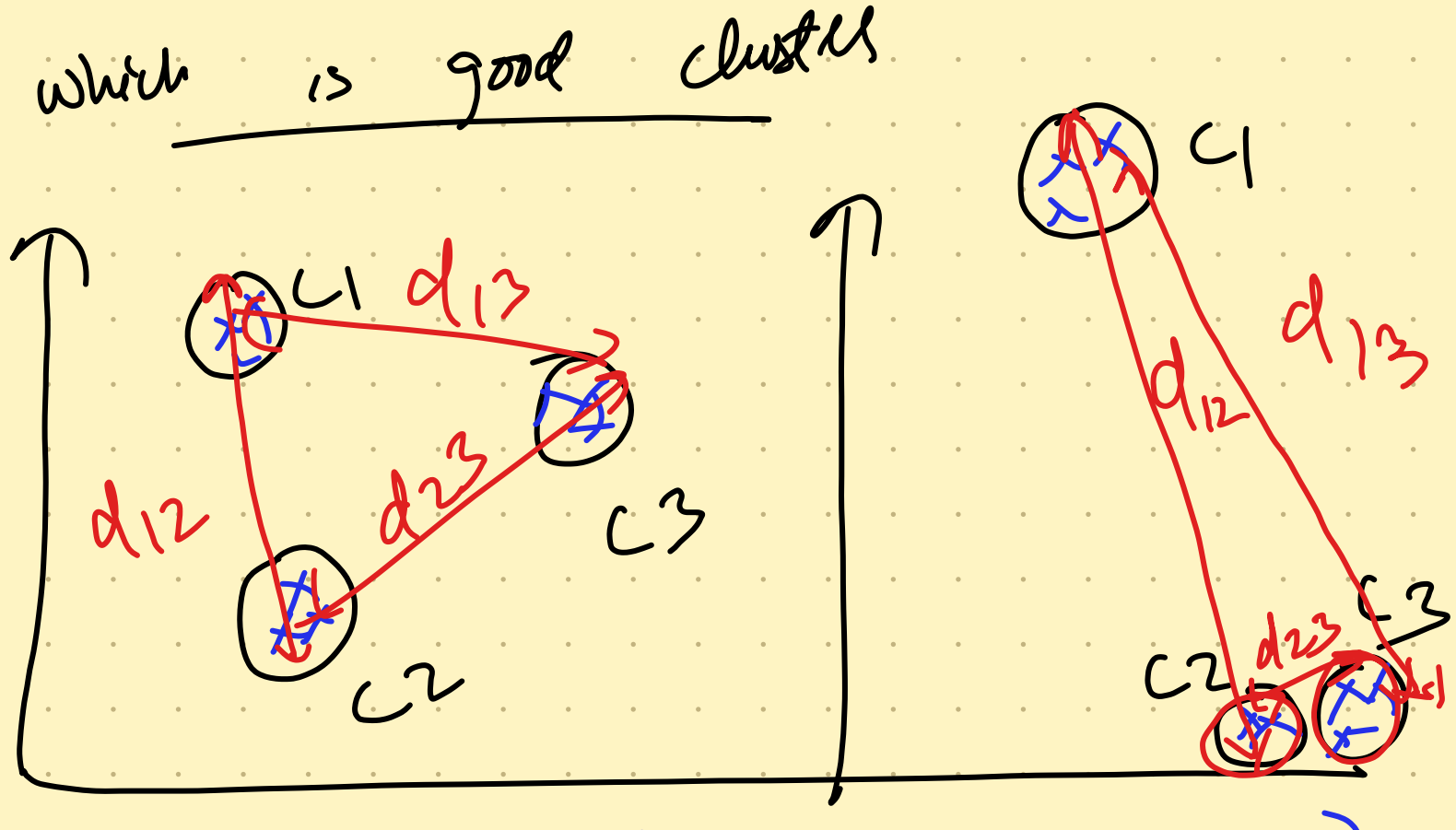


Intra cluster distance



$d'(i)$ = max distance b/w 2 points in C_i

Q which is good cluster



1 2

max $(d_{12}, d_{13} > d_{23})$ ⇒ A

$$\min(d_{12}, d_{13}, d_{23}) = \gamma$$

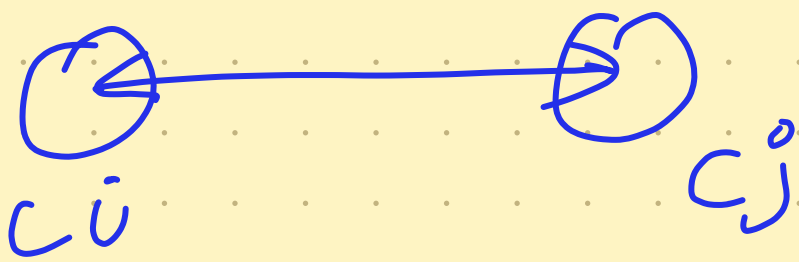
γ (case I) $>$ γ (case II)

Case I is a better cluster arrangement

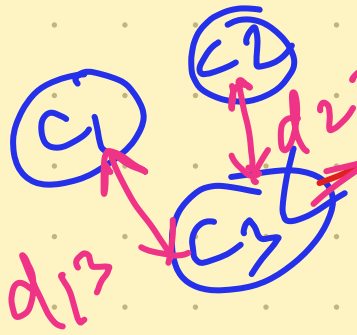
Intra Cluster Distance

Dunn Index $=$ $\left(\frac{\min(\text{Inter clusters dist})}{(\max)(\text{Intra cluster Dist})}\right)$

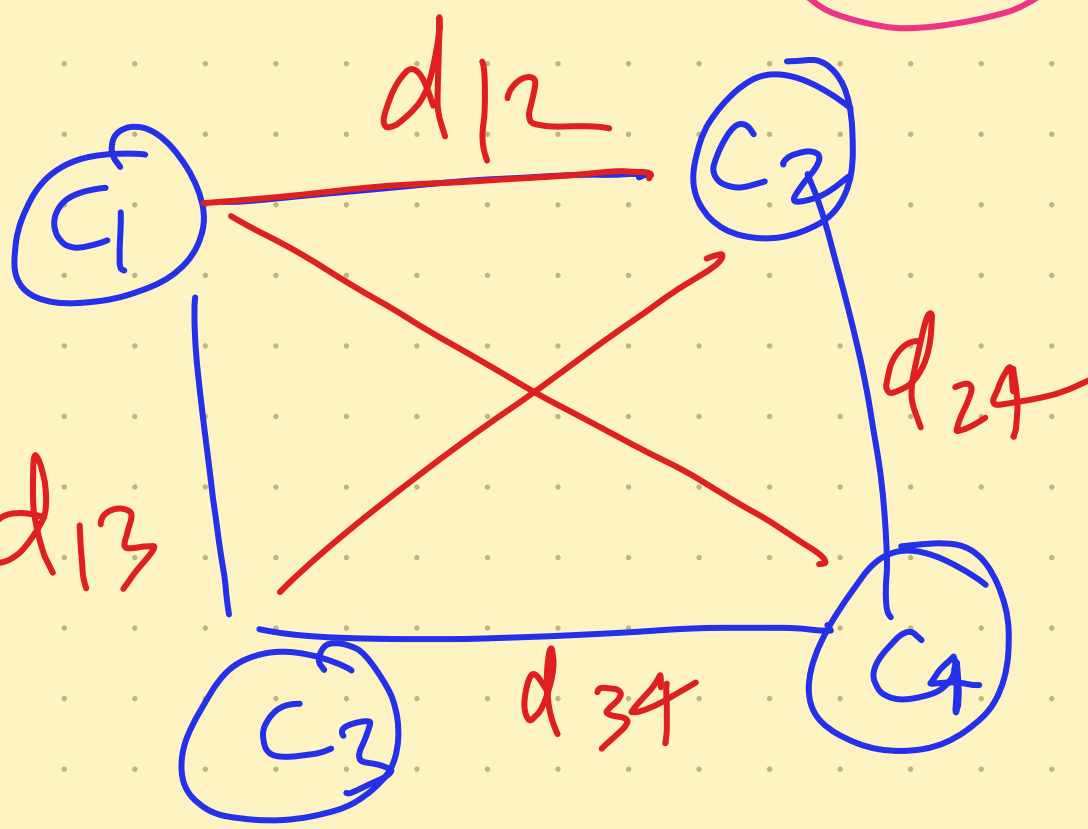
D.I $=$ $\frac{d(i,j)}{d'(i)}$ → i, j are from diff clusters



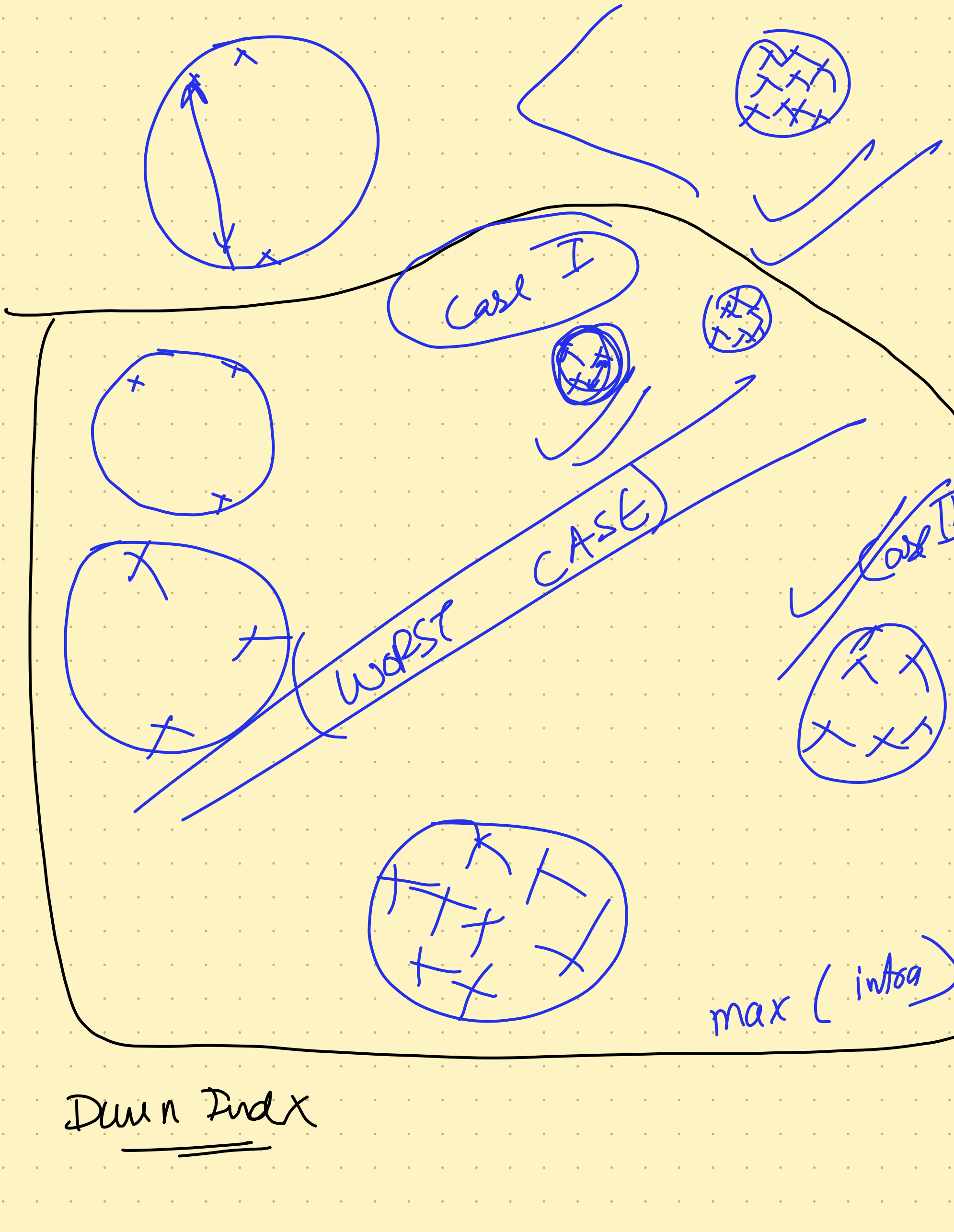
Case I



Case II



$\min\{\text{inter cluster distance}\}$ / $\{\max \text{ intra cluster}\}$

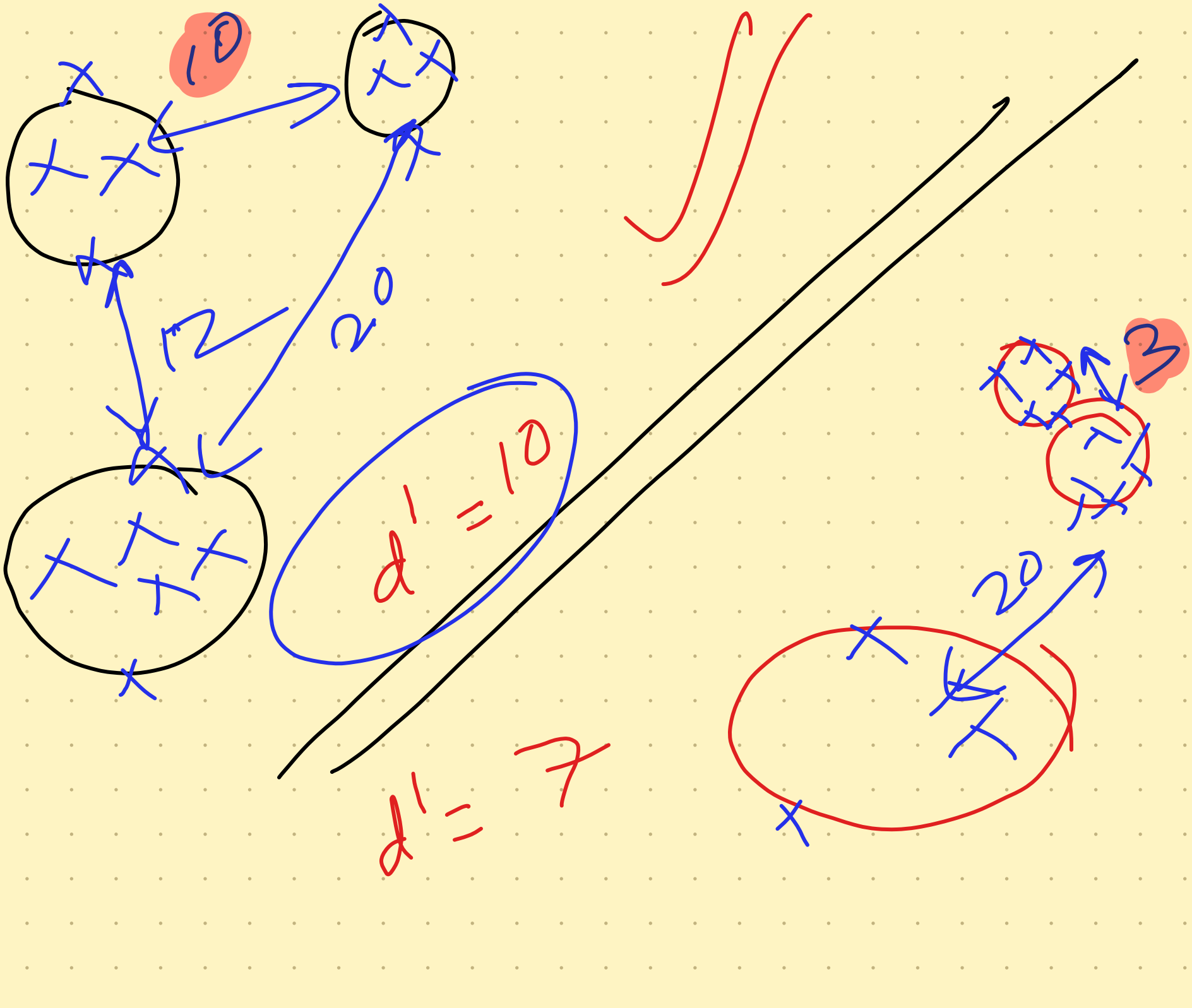


Dun Find

Dunn Index = $\frac{\min (\text{Inter})}{d' = \max (\text{Intra})}$ ↑ ↓

10

$\min (\text{Inter}) = 10 (\max \text{Intra})$



Dunn Index (I) = (0/10)

D.I(II) = (3/7)

Higher Dunn Index ⇒ Better Clustering Algo

O(n log n)

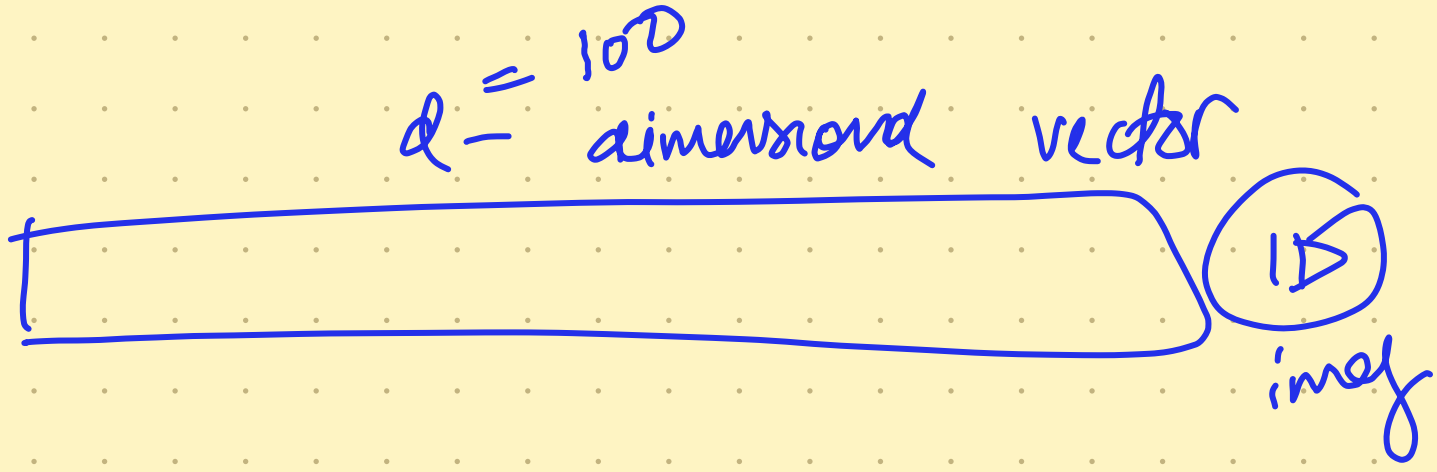
Applications

Jeans

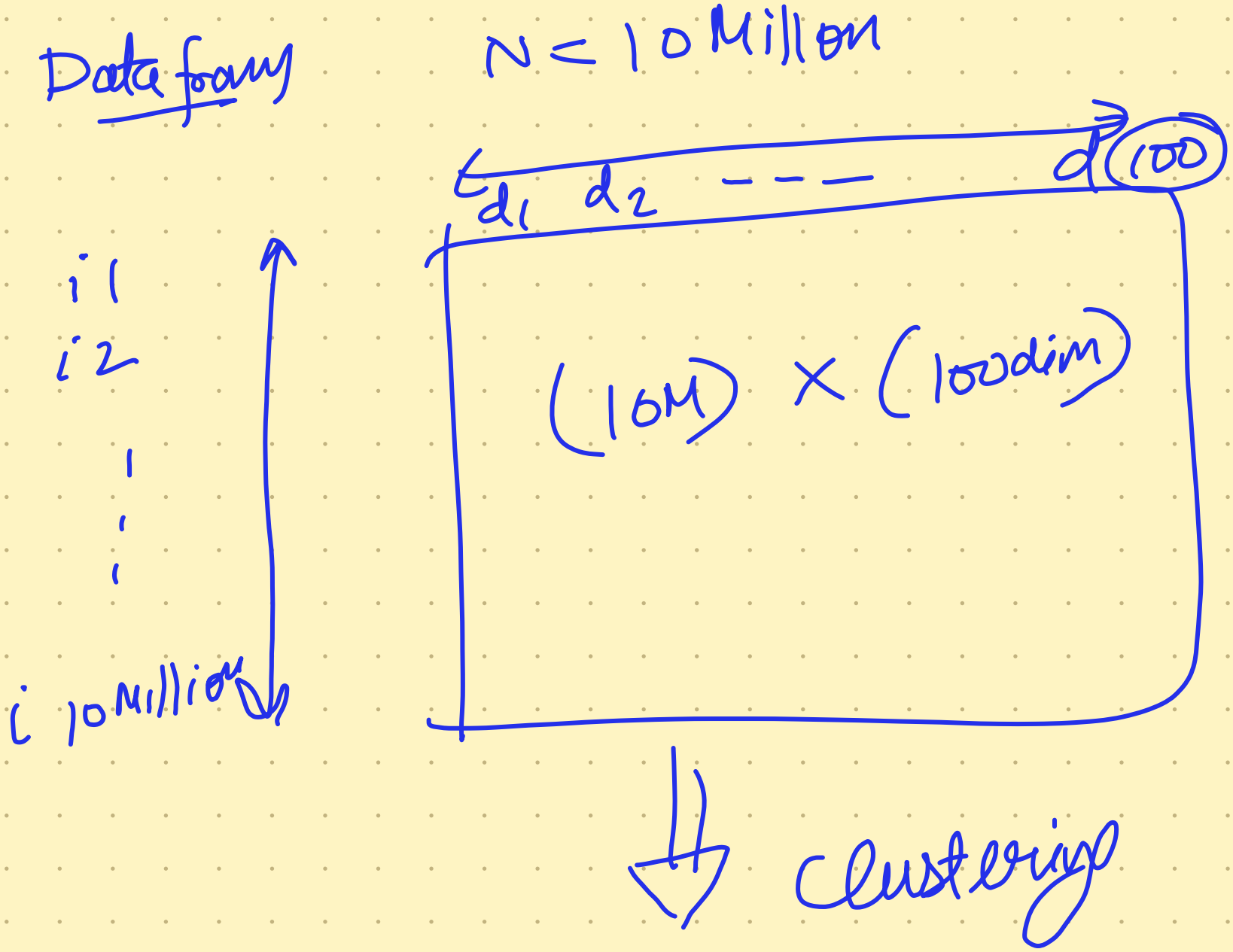
Top

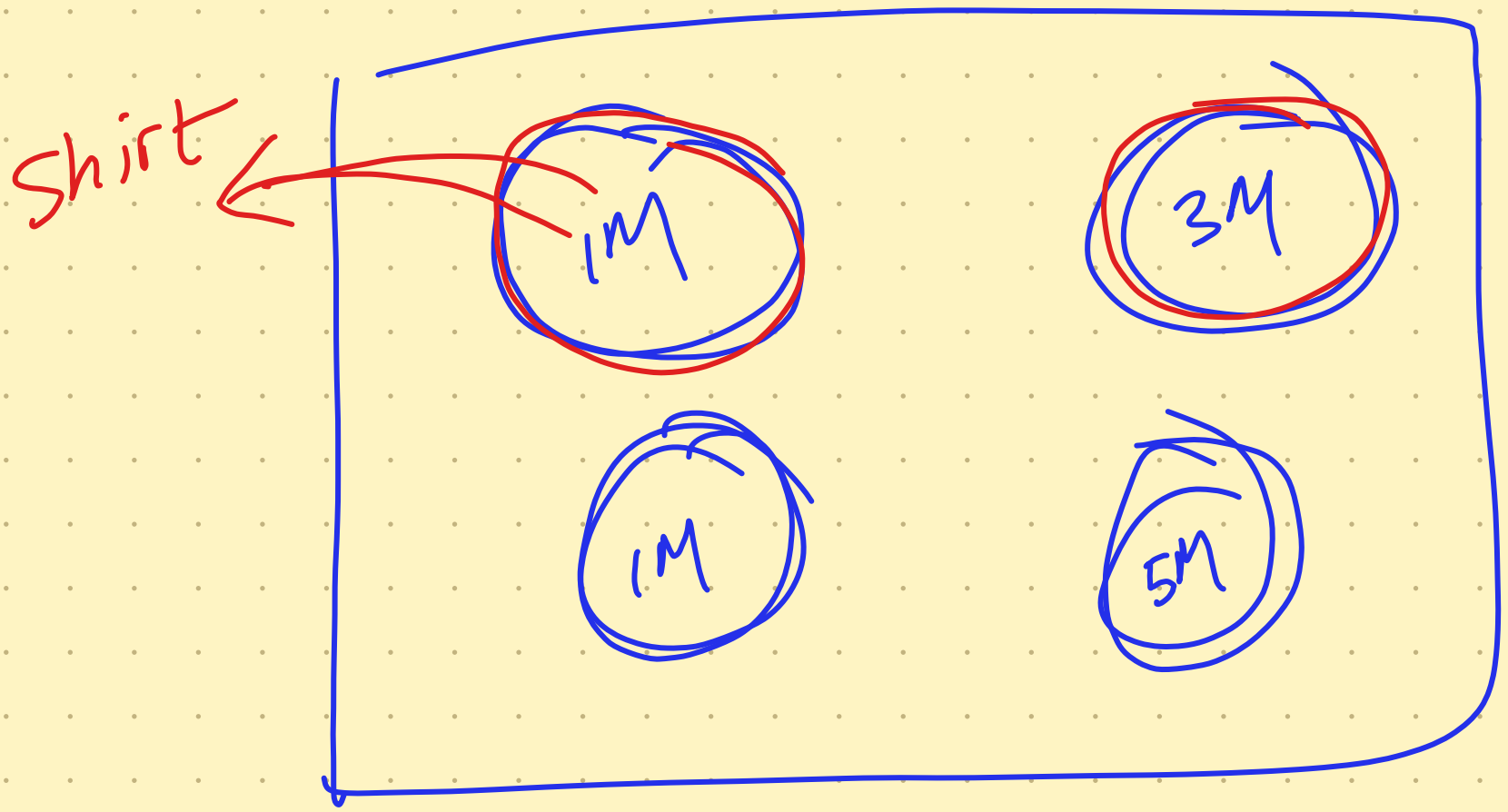
Shirt

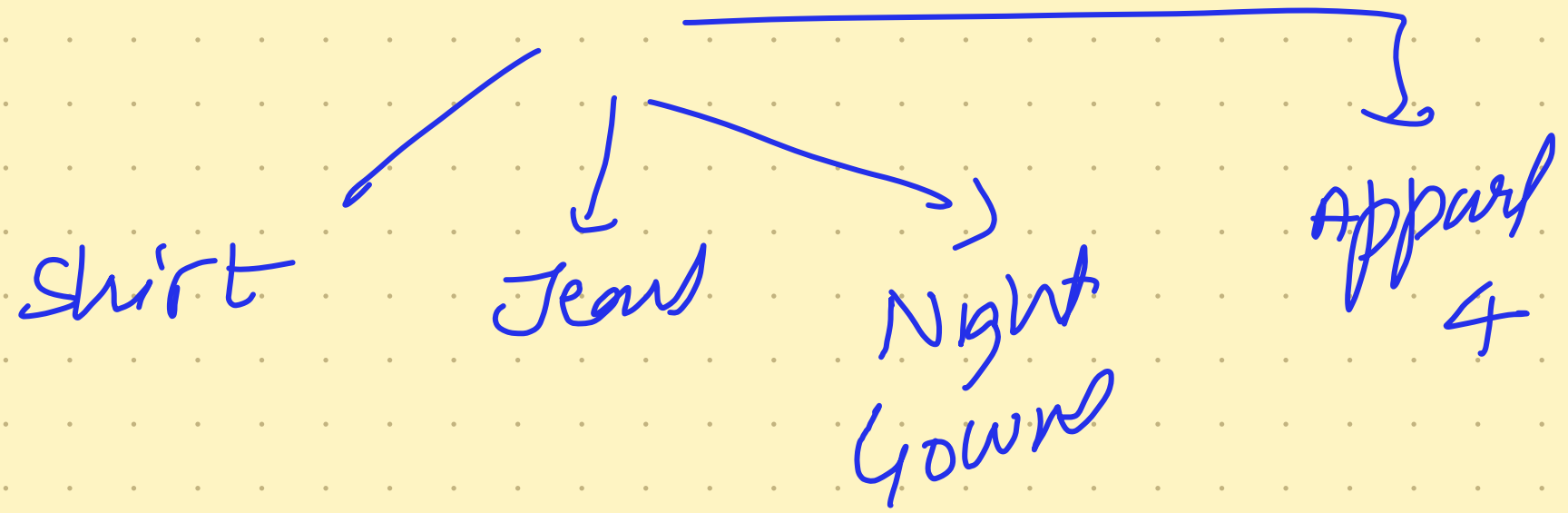
Night Gown



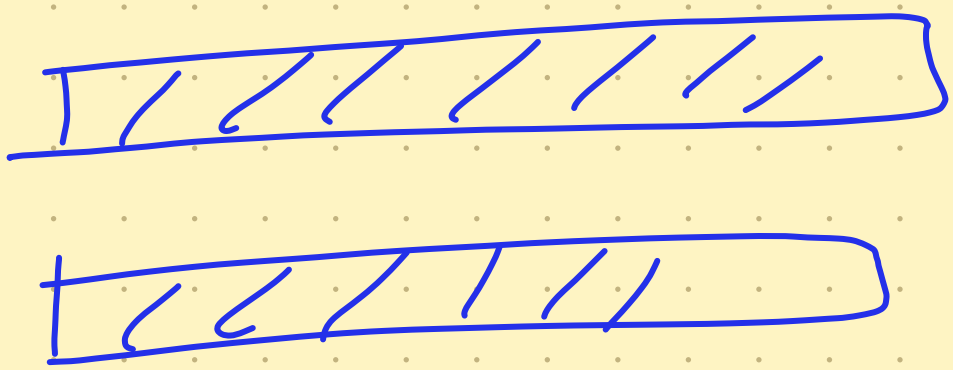
Data foamy







Document Labeling →

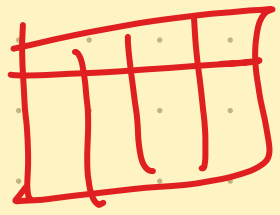


K Means Algorithm

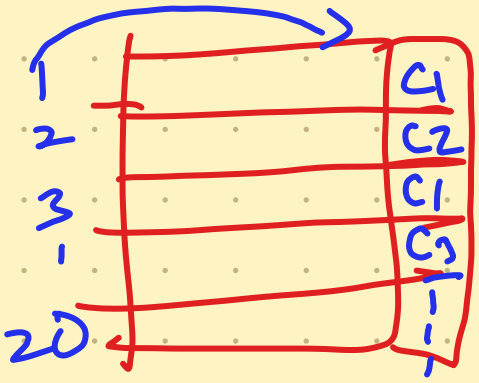
K → no of clusters

Clustering → Input (n) → d dim input

(#) K → clusters



K means = 3



output → K clusters

$C_1\ C_2 \dots C_K$

K means

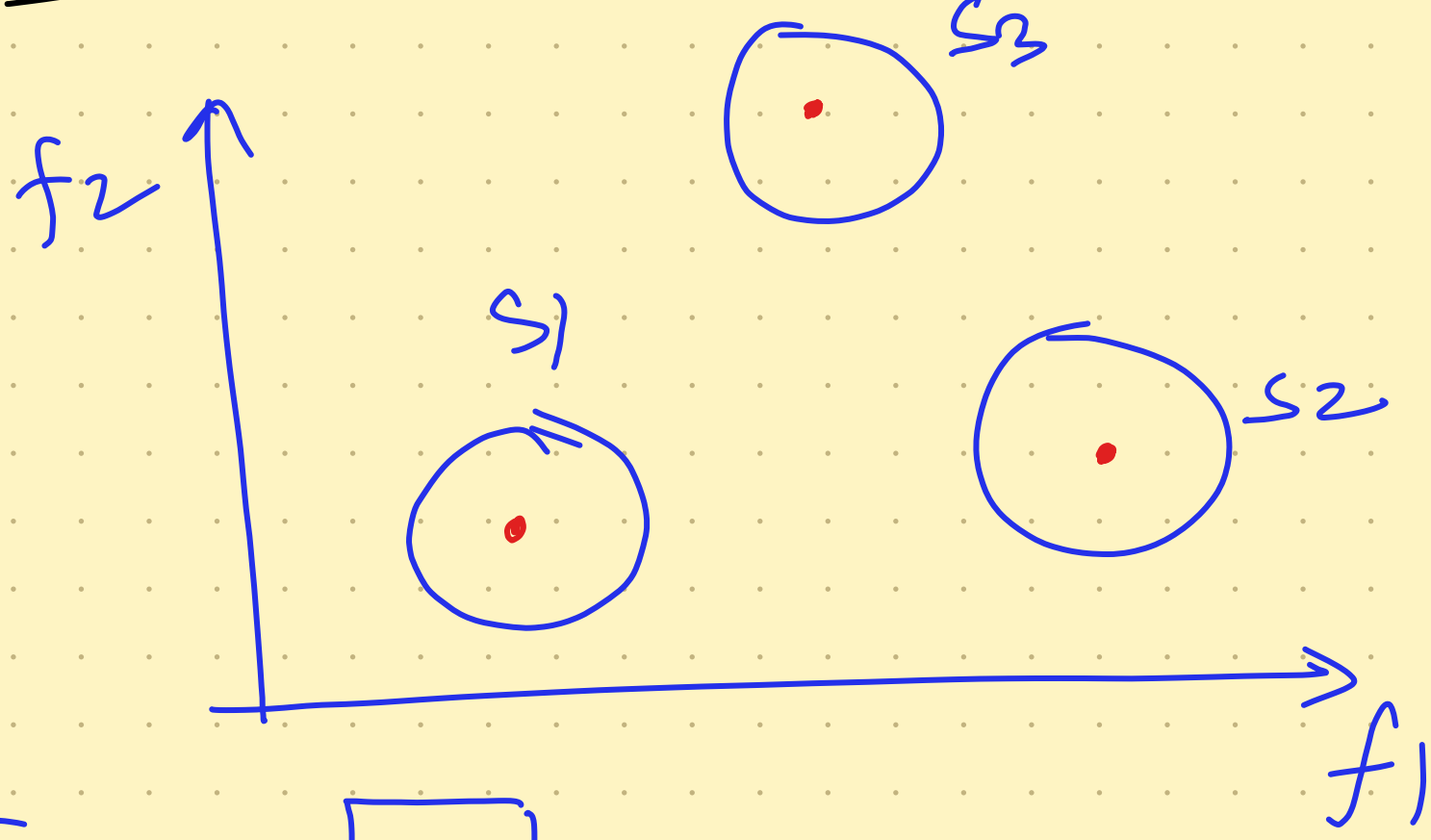
→ simple

→ optimz^n Algo

→ Populas

→ you need to define K before hand

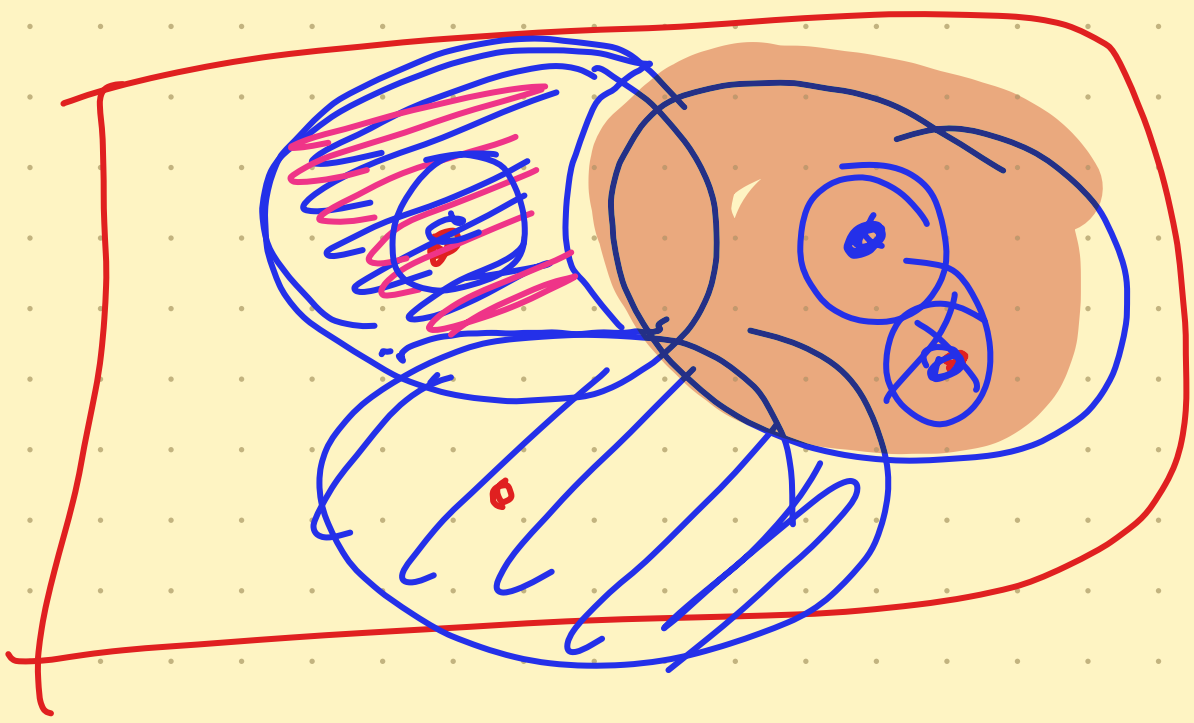
2 Dim dataset → 3 clusters (K=3)

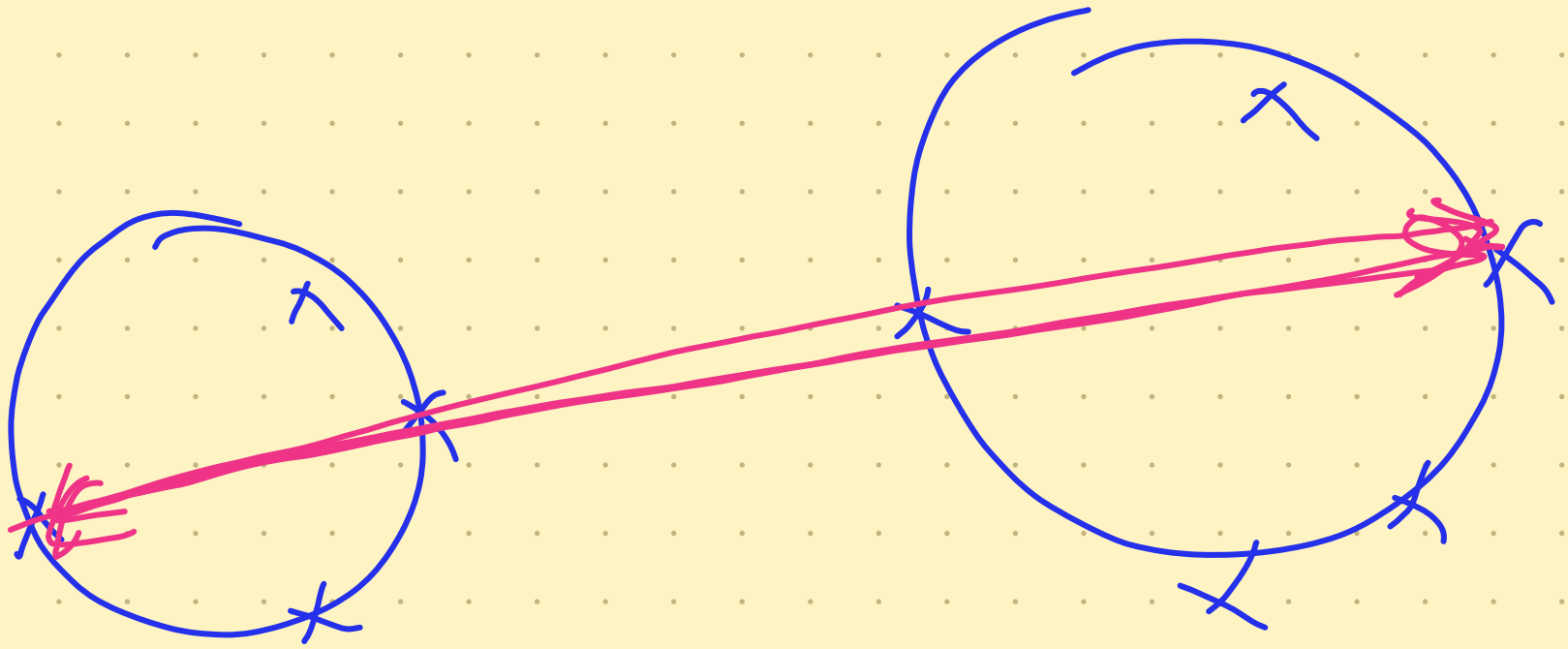


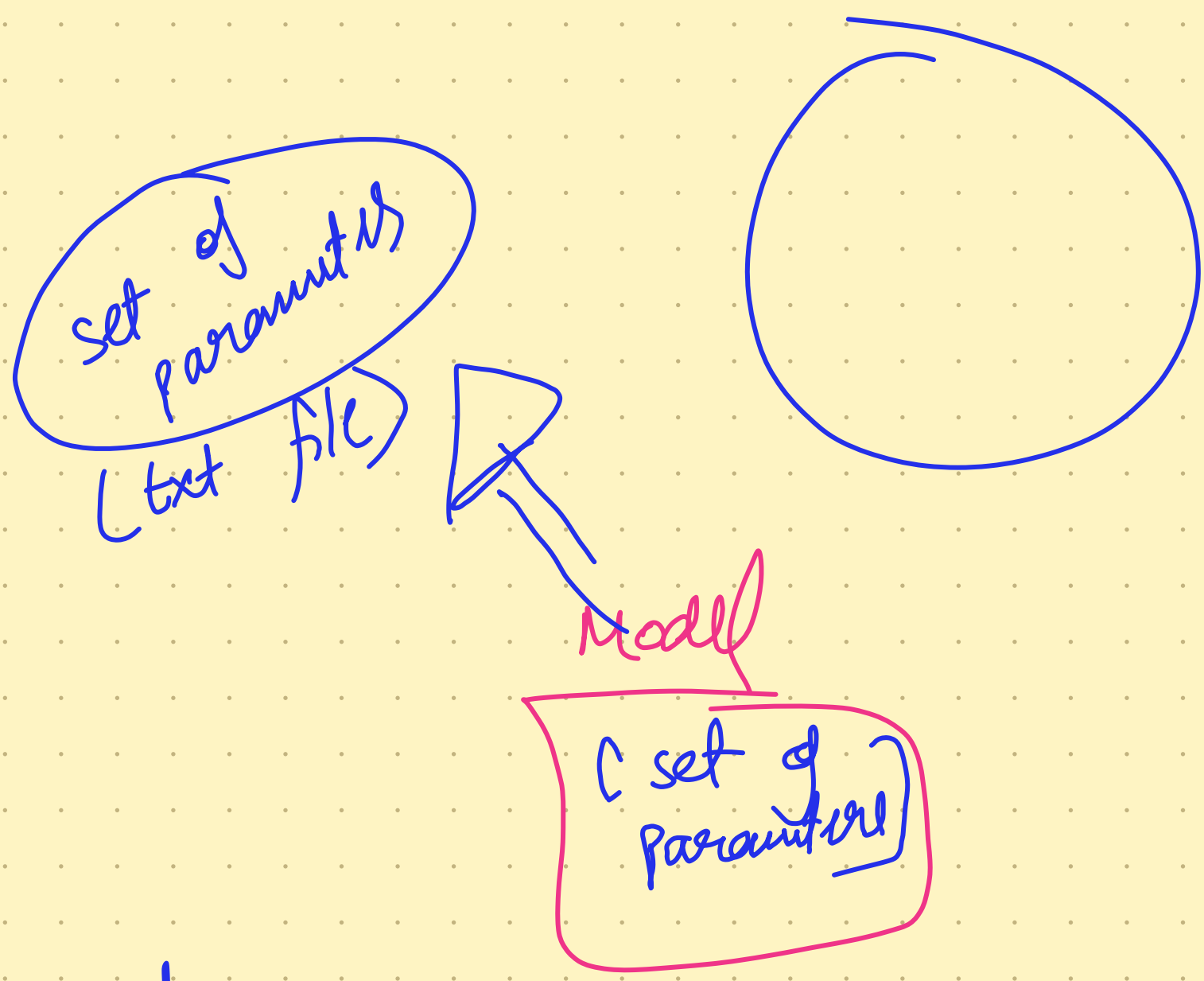
$S_i \rightarrow C_i$

$S_i \cap S_j = \phi$ and each x_i Gone S_j

Mean Data $= \left(\frac{x_1 + \cdots x_m}{m}\right) \in S_i$







Input → [○ ○ ○] → output

L.R → weights of model

Salary = w_1 (Age) + w_2 (Exp)

Age

saving

Exp

w_1 w_2

output

Case I

X

outer

loss

Case II

min

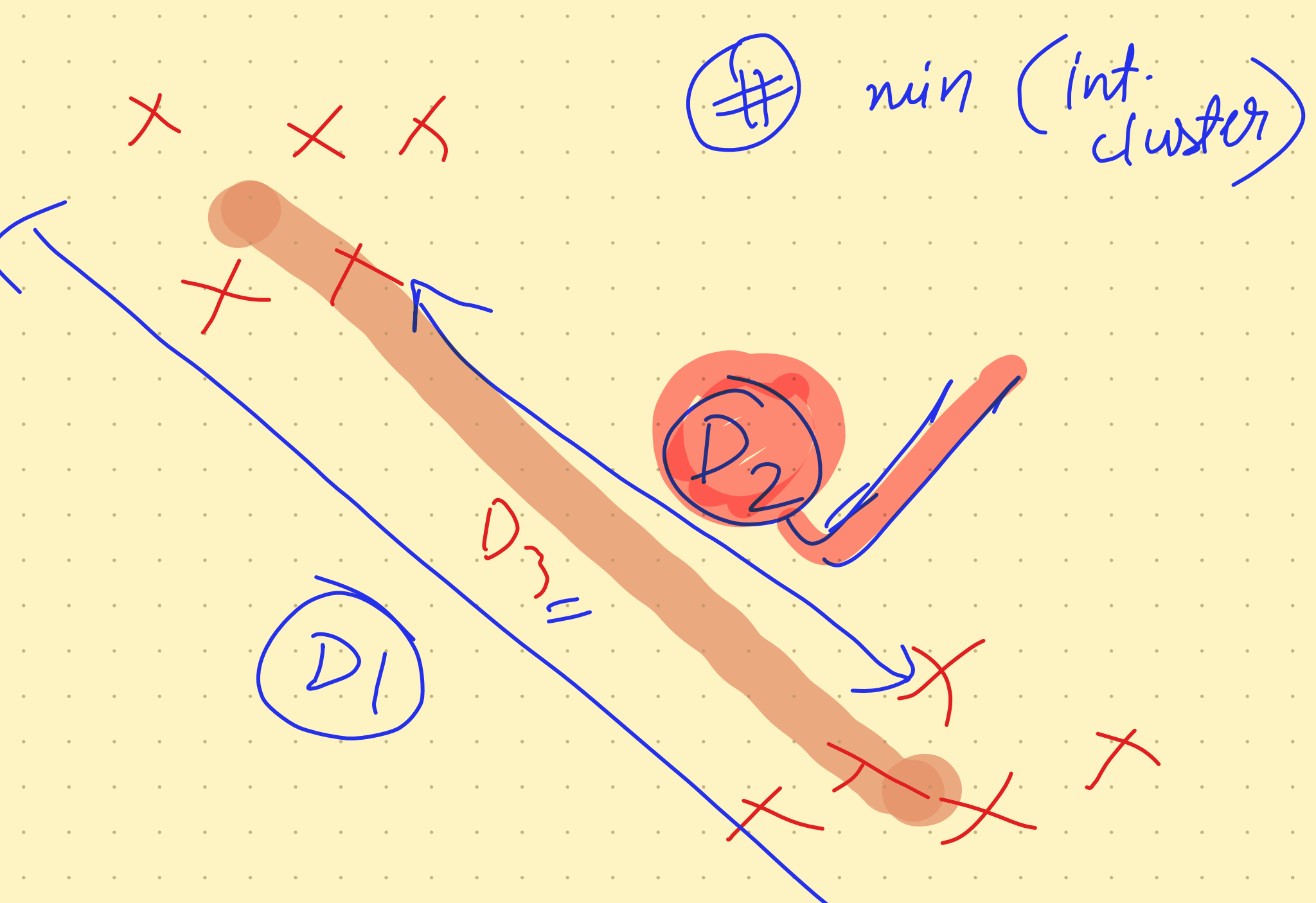
DBI

Good Cluster

Dunn Ind = min (Inter clus)

inter clusters (high)

intra clusters (low)



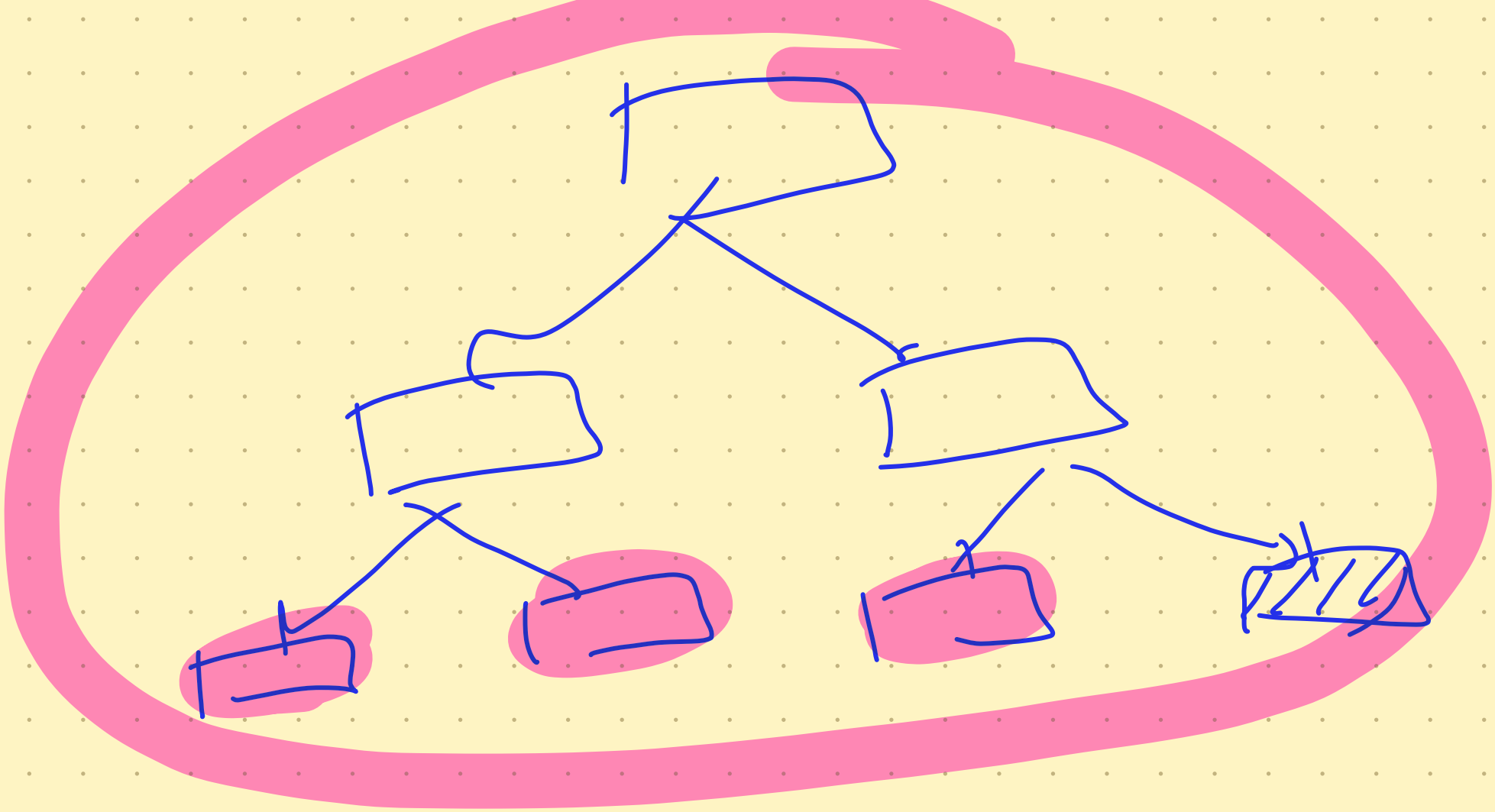
D1 and D2

training

D-tree Algo

test accuracy

Plot the Decision Tree



classifc^{n}

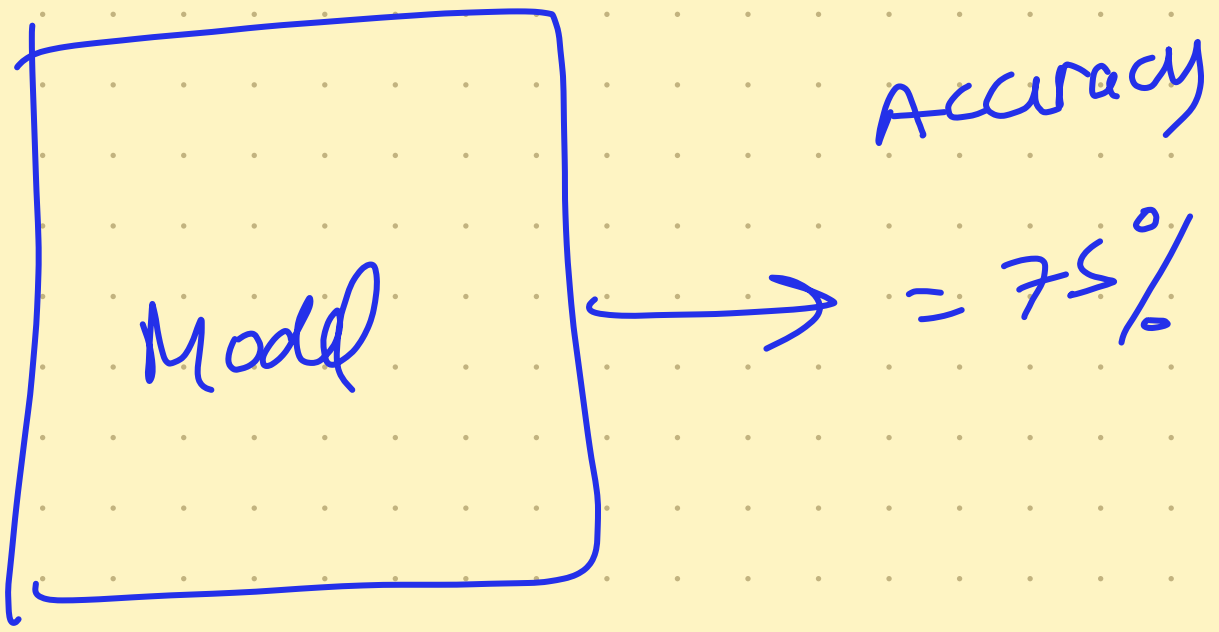
Precision
Recall
Confusion Matrix

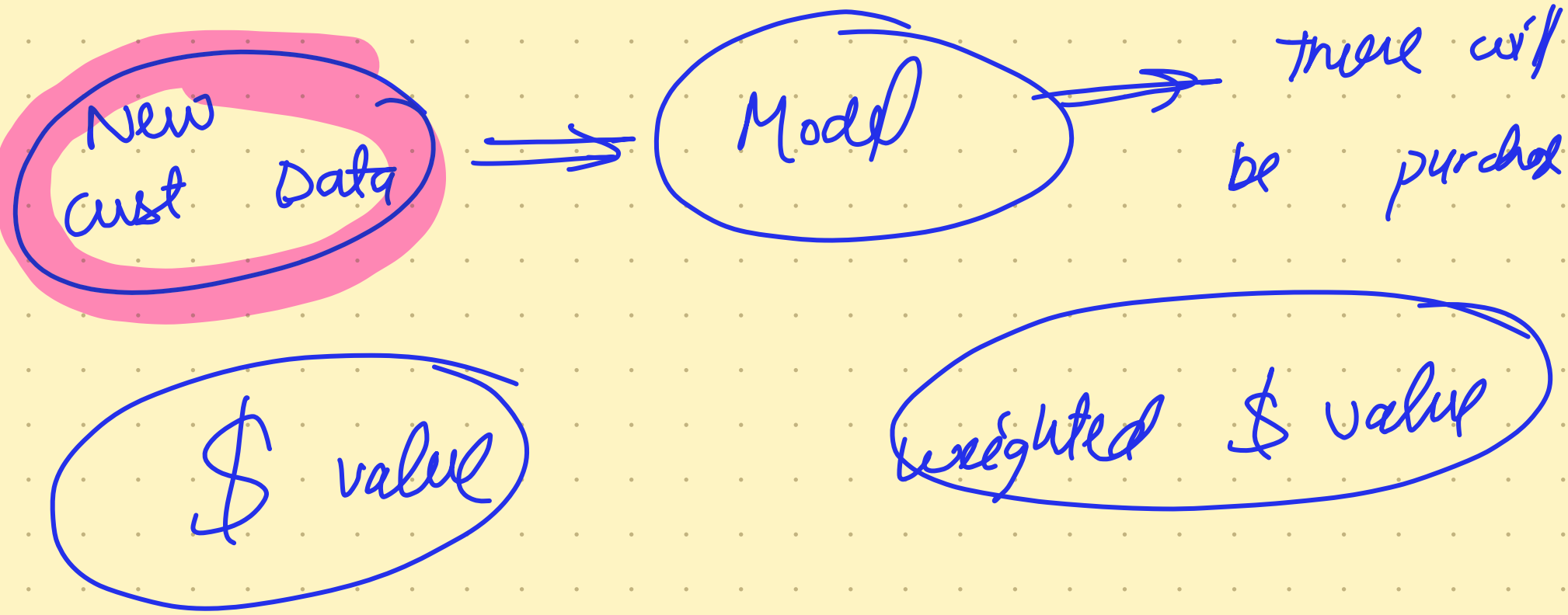
Purchas / Not → 75

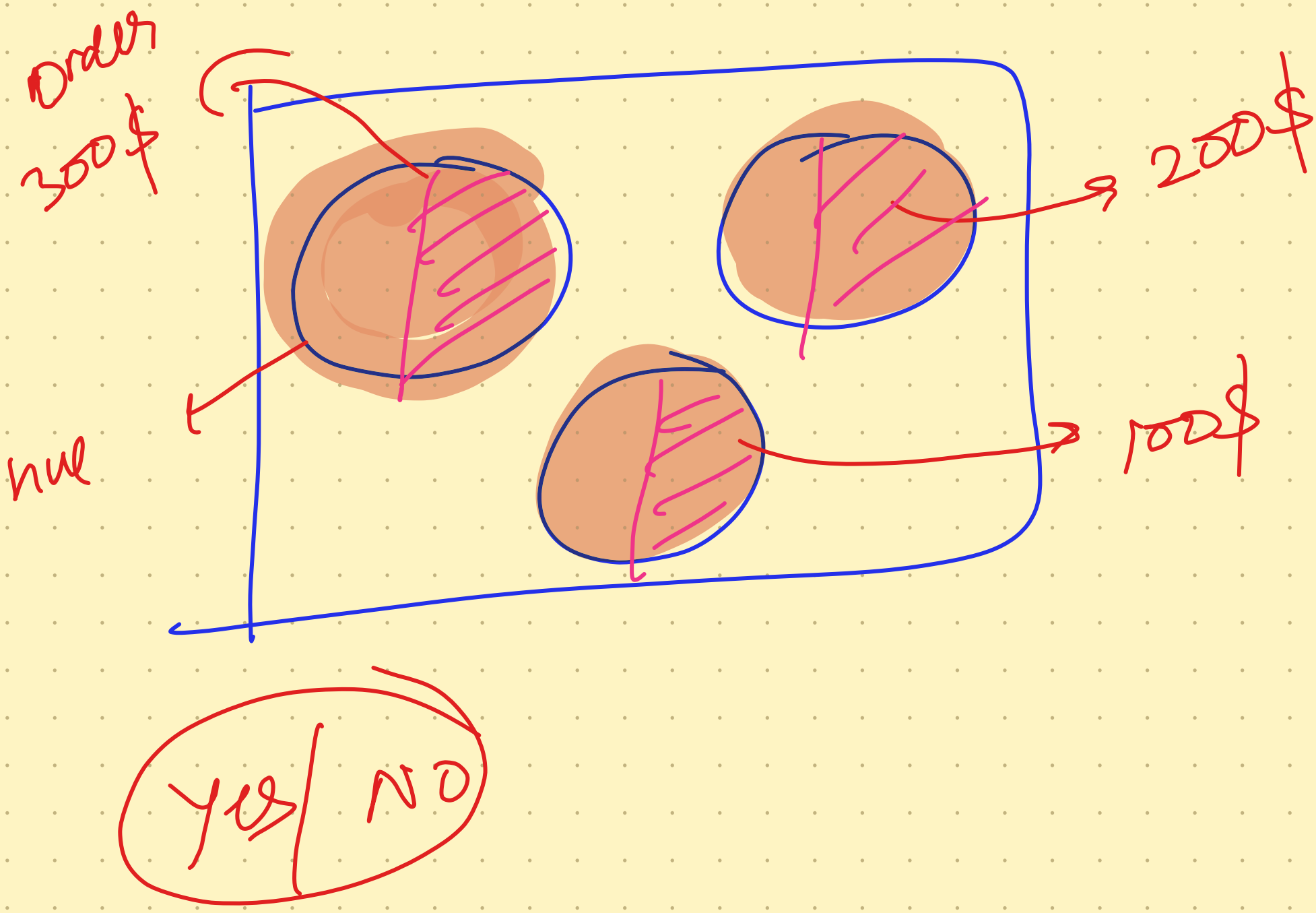
Accuracy = 75

← 1

Acur > 50%


Model
Accuracy
= 75%


New cust Data
Model
these will be purchd
$ value
weighted $ value
yes


Order
300$
200$
100$
null
YES/ NO